

Investigation and Implementation of Computer-Based Aids for the Speech Impaired

CATHERINE INEZ WATSON

A thesis presented for the degree of Doctor of Philosophy
in Electrical and Electronic Engineering at the
University of Canterbury, Christchurch, New Zealand.

May 1994

ABSTRACT

A computer-based visual speech aid has been developed to help those with speech defects. By transforming speech into visual patterns, the speech-impaired are shown visual representations of target speech features and of their attempts to reproduce these features. Real-time digital signal processing allows a speech therapist and client to use the speech aid interactively. The aid has been tested in the clinics of several therapists.

Two special tests have been designed to assess the speech aid's ability to separate acceptable and unacceptable speech, and to display crucial features of the speech. In addition, two new speech analysis algorithms have been developed for the aid. The first is a recognition algorithm for isolated fricative sounds of English. It is based on spectral classification. The second is an algorithm for reconstructing the shape of the vocal tract and it uses Wiener filtering to deconvolve the glottal pulse from the speech signal.

ACKNOWLEDGEMENTS

This doctorate has been a challenging journey; one I would not have been able to travel if it were not for the support, help, and encouragement of many people. I am extremely grateful to John Andreae and the late Richard Bates, two men whose quest for knowledge and excellence have been a great inspiration to me. Richard Bates provided the vision in the first half of my research and John Andreae taught me how to realize it. I would also like to thank John Andreae for the many many hours he has spent proof reading my thesis and coaxing logical statements out of me. I am also very grateful for the invaluable advice and help given to me by Bill Kennedy and Margaret MacLagan, and for the proof reading they have done.

I have thoroughly enjoyed my association with the Speech Group. Our weekly meetings have resulted in many interesting discussions and ideas. I will always appreciate the support and encouragement the group provided; so to Tracy Clark, Andrew Elder, John Kirkland, Brenda Satherley, William Thorpe, and Nicola Woods, thank you. I am very grateful to the Canterbury Speech Therapy Community for giving me their time, evaluating the CASTT (which has been invaluable), and for their encouragement. In particular I would like to thank Evelyn Teriss for her unfailing enthusiasm for the CASTT.

This study would not have been possible if it were not for the financial assistance of the 1988 Tait Electronics Ltd Research Scholarship (Postgraduate), the New Zealand Federation of University Women Postgraduate Fellowship (1989), the Department of Science and Industrial Research Postgraduate Women's Study Award (1989), the Royal Society of New Zealand - Canterbury Branch Travel Scholarship (1990) and Electronic Design Associates, Costa Mesa, U.S.A..

My stay here in the Department of Electrical and Electronic Engineering has been made particularly enjoyable by the many friends I have made both in and outside the department. I would particularly like to thank the past and present inmates of R6, Pam, Denise, Sarah, Viv and Wayne for tolerating my absent mindedness, humouring me when my car had run out of petrol yet again, bringing in my washing, providing a kind word when needed, and always being there. Finally I would like to thank my family; words are not adequate to express my gratitude to you all. Megan, Geoffrey, and Grandma: you have always been unceasingly encouraging. Mum and Dad: you have always been a tower of strength, your faith in me has provided me with the tenacity and will to complete this work.

CONTENTS

| | |
|--|-------------|
| ABSTRACT | iii |
| ACKNOWLEDGEMENTS | v |
| PREFACE | xiii |
| CHAPTER 1 ASPECTS OF SPEECH | 1 |
| 1.1 Introduction | 1 |
| 1.2 Anatomy and Physiology of Speech Producing Mechanisms | 2 |
| 1.2.1 The Respiration Structure and its Function | 2 |
| 1.2.2 The Laryngeal Structure and Phonation | 4 |
| 1.2.3 The Supralaryngeal Structures and Articulation | 6 |
| 1.2.4 The Innervation System and Speech Control | 7 |
| 1.3 The Nature of Speech | 8 |
| 1.3.1 Segmental Aspects Of Speech | 8 |
| 1.3.1.1 Phonological Structure | 9 |
| 1.3.1.2 Articulation | 11 |
| 1.3.2 Suprasegmental Aspects Of Speech | 14 |
| 1.4 The Acoustic Features of Speech | 14 |
| 1.5 Disorders of the Spoken Language and Speech and Language Therapy | 17 |
| CHAPTER 2 VISUAL SPEECH AIDS | 21 |
| 2.1 Introduction | 21 |
| 2.2 Non-Computer-Based Visual Speech Aids | 22 |
| 2.3 Computer-Based Visual Speech Aids | 23 |
| 2.3.1 Pitch Correctors | 24 |
| 2.3.2 Loudness Correctors | 25 |
| 2.3.3 Sustained Phonation and Voicing Correctors | 26 |
| 2.3.4 Articulation Correctors | 27 |
| 2.3.5 Other Speech Correctors | 29 |
| 2.3.6 The Hardware Components and User-Interface Software | 31 |
| 2.4 Conclusion | 32 |
| CHAPTER 3 THE COMPUTER AIDED SPEECH THERAPY TOOL (CASTT) | 35 |
| 3.1 Introduction | 35 |
| 3.2 The CASTT | 35 |
| 3.2.1 The Voice Pitch Tracker Module | 36 |

| | | |
|------------------|--|-----------|
| 3.2.2 | Loudness Monitor | 38 |
| 3.2.3 | The Concurrent Loudness And Pitch Module | 38 |
| 3.2.4 | The Spectrogram Module | 39 |
| 3.2.5 | The Fricative Monitor | 40 |
| 3.2.6 | The Vocal Tract Shape Module | 40 |
| 3.2.7 | The Sustained Phonation Module | 41 |
| 3.2.8 | The CASTT Shell Software Package | 42 |
| 3.2.9 | The Hardware Of The CASTT | 42 |
| CHAPTER 4 | SIGNAL PROCESSING FOR THE CASTT | 45 |
| 4.1 | Digitized Speech | 45 |
| 4.2 | The Hardware And Software Of The CASTT | 46 |
| 4.2.1 | Software | 47 |
| 4.2.2 | Programming | 49 |
| 4.3 | Loudness Analysis | 50 |
| 4.3.1 | The TMS32010 Root Mean Square Algorithm | 50 |
| 4.3.2 | The Modules Which Use The TMS32010 RMS algorithm | 50 |
| 4.4 | Pitch Analysis | 51 |
| 4.4.1 | The TMS32010 Pitch Algorithm | 54 |
| 4.4.1.1 | Peak Detection | 54 |
| 4.4.1.2 | Pitch Extracting | 56 |
| 4.4.2 | The Modules Which Use The TMS32010 Pitch Algorithm | 59 |
| 4.5 | Spectral Analysis | 59 |
| 4.5.1 | Fourier Analysis | 61 |
| 4.5.1.1 | The Discrete Fourier Transform | 62 |
| 4.5.2 | Windowing the Time Domain Signal | 62 |
| 4.5.3 | The TMS32010 Spectrogram Algorithm | 63 |
| 4.5.3.1 | Calculating The Magnitude Spectrum | 63 |
| 4.5.3.2 | Assigning Colour and Display Coordinates To The Data Points Of The Spectrogram | 64 |
| 4.5.4 | The Modules Which Used The TMS32010 Spectrogram Algorithm | 65 |
| 4.6 | Zero-Crossing Analysis | 65 |
| 4.6.1 | The TMS32010 Zero-Crossing Algorithm | 65 |
| 4.7 | The Vocal Tract Reconstruction | 67 |
| 4.7.1 | The Vocal Tract Area Reconstruction Algorithm | 68 |
| 4.7.2 | Calculating the Vocal Tract Shape | 68 |
| 4.7.2.1 | Calculating The Front Of The Vocal Tract Co-ordinates | 70 |
| 4.7.2.2 | Calculating The Jaw Co-ordinates | 71 |
| 4.8 | Lissajous Figures - A Potential Speech Analysis Module for the CASTT | 72 |
| 4.8.1 | Phase Shifting In the Frequency Domain | 74 |
| 4.8.2 | The Lissajous Figure Algorithm | 74 |
| CHAPTER 5 | EVALUATION OF CASTT BY SPEECH THERAPISTS | 77 |
| 5.1 | The Short-Term Evaluations Of The CASTT | 78 |
| 5.1.1 | The First Assessment Period Of The CASTT | 79 |

| | | |
|---|--|------------|
| 5.1.1.1 | The Consequences Of The First Assessment Period | 79 |
| 5.1.2 | The Second Assessment Period | 83 |
| 5.1.2.1 | The Consequences Of The Second Assessment Period | 83 |
| 5.2 | The Long-Term Evaluations of the CASTT | 84 |
| 5.2.1 | Three Therapists' Comments about the CASTT | 85 |
| 5.3 | The Consequences Of The Therapists' Evaluations | 87 |
| CHAPTER 6 AN EVALUATION OF THE CASTT VISUAL DISPLAYS | | 91 |
| 6.1 | Introduction | 91 |
| 6.2 | Assessment Of Visual Displays Of Computer-Based Speech Therapy Aids | 92 |
| 6.2.1 | The Short Test Of Elementary Error Discrimination | 93 |
| 6.3 | The Preliminary Test Of The CASTT's Visual Displays | 93 |
| 6.3.1 | The Modification Of The Speech List Which Represents The Elementary Errors | 95 |
| 6.3.2 | Obtaining The Pre-Recorded Speech For The Preliminary VDT | 96 |
| 6.3.3 | Preparing The Visual Displays For The Preliminary VDT | 97 |
| 6.3.4 | The Participants Of The Preliminary VDT | 98 |
| 6.3.5 | The Results Of The Preliminary VDT | 98 |
| 6.3.6 | Discussion Of The Preliminary VDT Results | 99 |
| 6.4 | The Visual Display Test (VDT) | 100 |
| 6.4.1 | The Preparation Of The VDT | 100 |
| 6.4.1.1 | Obtaining The Pre-Recorded Speech For The VDT | 100 |
| 6.4.1.2 | Preparing The Visual Displays For The VDT | 101 |
| 6.4.1.3 | The Elementary Errors For Which The Display Types Were Tested | 101 |
| 6.4.2 | The Presentation Of The VDT | 101 |
| 6.4.3 | The Participants Of The VDT | 103 |
| 6.4.4 | Remedial Potential | 103 |
| 6.4.5 | The Results Of The VDT | 106 |
| 6.4.5.1 | The Remedial Potential Of Each Of The Visual Display Types | 107 |
| 6.5 | The Visual Display Test Part II | 115 |
| 6.5.1 | The Time-Plots | 117 |
| 6.5.1.1 | The Displays Of Loudness Contours | 117 |
| 6.5.1.2 | The Displays Of Pitch Contours | 132 |
| 6.5.1.3 | The Spectral Content Display | 140 |
| 6.5.2 | The Current-Value-Plots | 151 |
| 6.5.2.1 | The Vocal Tract Shape Display | 152 |
| 6.5.2.2 | The Lissajous Figure Displays | 154 |
| 6.5.2.3 | The Fricative Monitor Display | 157 |
| 6.5.3 | Discussion | 160 |
| CHAPTER 7 COMPUTER RECOGNITION OF FRICATIVE SOUNDS | | 163 |
| 7.1 | Introduction | 163 |

| | | |
|---|--|------------|
| 7.2 | Phonetics of Fricatives | 164 |
| 7.3 | Time Domain Fricative Classification | 164 |
| 7.3.1 | Classification due to Waveform Intensity | 164 |
| 7.3.2 | Fricative Duration | 165 |
| 7.3.3 | Classification According To The Zero-Crossing Rate Of The Waveform | 166 |
| 7.4 | Frequency Domain Fricative Classification | 166 |
| 7.4.1 | Fricative Spectral Shapes | 168 |
| 7.5 | Automatic Fricative Sorting Algorithms | 170 |
| 7.5.1 | The McKinnon and Lee Zero-Crossing Rate Based Algorithm | 170 |
| 7.5.2 | Algorithms Based On Spectral Shape | 171 |
| 7.5.2.1 | The Hughes and Halle Algorithm | 171 |
| 7.5.2.2 | The Jassem Algorithm | 172 |
| 7.5.2.3 | Jassem's Envelope Polynomial | 173 |
| 7.5.2.4 | Molho's Spectral Envelope Descriptors | 173 |
| 7.5.2.5 | The Success Rates Of Several Fricative Sorting Algorithms | 174 |
| 7.6 | A Preliminary Investigation Into A Fricative Recognition Algorithm | 174 |
| 7.6.0.6 | Obtaining The Fricative Spectrum | 176 |
| 7.6.1 | The Algorithm | 178 |
| 7.6.2 | Labiodental/Dental, Alveolar and Palatal Distinction | 178 |
| 7.6.2.1 | Labiodental/Dental Distinction | 178 |
| 7.6.2.2 | The Voiced/Unvoiced Distinction | 179 |
| 7.6.3 | Validating The Proposed Sorting Algorithm | 181 |
| 7.6.3.1 | Obtaining The Fricative Data | 181 |
| 7.6.3.2 | The Results | 181 |
| 7.7 | The Implications of the Fricative Recognition Algorithm For The CASTT | 183 |
| CHAPTER 8 VOCAL TRACT SHAPE RECONSTRUCTION | | 185 |
| 8.1 | A Brief Introduction To Some Past Reconstruction Techniques | 185 |
| 8.2 | The Acoustic Tube Model Of The Vocal Tract | 186 |
| 8.2.1 | Relating the Vocal Tract Shape to the Acoustic Signals | 186 |
| 8.2.2 | Relating The Reflection Coefficients Of The Concatenated Tube Model To The Vocal Tract Shape | 188 |
| 8.2.2.1 | Solutions To The Conservation Equations | 188 |
| 8.2.2.2 | Relating The Pressure and Bulk Flow Wave Components | 189 |
| 8.2.2.3 | Behaviour Of Planar Waves At Boundaries Between Two Uniform Tubes | 189 |
| 8.2.2.4 | A Gross Assumption Of The Concatenated Acoustic Tube Model Of The Vocal Tract | 191 |
| 8.3 | Vocal Tract Shape Reconstruction Methods | 192 |
| 8.3.1 | The Impedance Tube Method Of Vocal Tract Reconstruction | 192 |
| 8.3.1.1 | Sondhi's Method | 192 |
| 8.3.1.2 | The Laval University Group Method | 194 |
| 8.3.2 | The Direct Estimation Method | 196 |
| 8.3.2.1 | The Linear Model Of Speech Production | 197 |

| | | |
|-------------------|--|------------|
| 8.3.2.2 | The Linear Prediction Model of Speech | 198 |
| 8.3.2.3 | Relating The Linear Prediction Model To The Acoustic Tube Model of Speech | 201 |
| 8.3.2.4 | Defining The Boundary Condition On The Acoustic Tube Model When Using The Direct Estimation Method | 204 |
| 8.3.2.5 | The Direct Estimation Vocal Tract Shape Reconstruction Method | 205 |
| 8.3.3 | Vocal Tract Shape Reconstruction From Acoustic Measures At Two Points | 206 |
| 8.3.4 | Discussion Of Vocal Tract Reconstruction Methods | 206 |
| 8.3.4.1 | The Problems Associated With The Direct Estimation Method | 207 |
| 8.4 | Vocal Tract Reconstruction By Inverse Filtering - Preliminary Investigation | 209 |
| 8.4.1 | The Wiener Filter | 209 |
| 8.4.2 | The Preparation For Vocal Tract Reconstruction By Inverse Filtering | 210 |
| 8.4.2.1 | The Guillemin Distance Measure | 211 |
| 8.4.2.2 | Making The Wiener Filter | 212 |
| 8.4.2.3 | The Importance Of Zero Extension | 213 |
| 8.4.3 | The Effect Of The Excitation Function On The Accuracy Of The Vocal Tract Reconstruction | 215 |
| 8.4.4 | The Effect Of Zero-Extension On The Reconstruction Process | 215 |
| 8.4.5 | The Effect Of The Number Of Pitch Periods On The Area Reconstruction | 216 |
| 8.4.6 | Discussion About The New Method Of Glottal Pulse Removal | 219 |
| CHAPTER 9 | CONCLUSIONS | 221 |
| APPENDIX A | QUESTIONNAIRES | 225 |
| A.1 | The Questionnaire For The First Assessment Period | 225 |
| A.2 | The Questionnaire For The Second Assessment Period | 225 |
| APPENDIX B | THE VISUAL DISPLAY TEST | 231 |
| B.1 | Data For Preliminary VDT | 231 |
| B.2 | Probabilities used in the VDT | 232 |
| REFERENCES | | 235 |

PREFACE

It is not often one gets a chance in life to practise what you preach and to become immersed in a subject that has always been a fascination. When the late Richard Bates offered me an opportunity to join the Speech Therapy aid project I was able to take just that chance. I have always believed that the directions of engineering development should be motivated by the needs of the community. Throughout the course of my research, through my interactions with the speech therapy community, I have endeavoured to keep the requirements of speech therapy and the needs of the therapist and client upmost in my mind. In addition, the sounds of speech have always fascinated me. I have always felt that Professor Henry Higgins in 'My Fair Lady' had rather a good job. Whilst my research has not yet resulted in me travelling to far flung places of the world to study speech, it certainly has been just as interesting.

During the course of my research, in addition to this thesis, I have written the following papers on the Speech Therapy Aid:

C.I. Watson, T.M. Clark, A.G. Elder and C.W. Thorpe (1988), "Multifarious Real-Time Speech Processing Applications", In Proc. NELCON, (*New Zealand National Electronics Conference*), Christchurch, September, p65-70.

C.I. Watson and R.H.T. Bates (1989), "Computerised Accessory For Speech Therapy", *New Zealand Medical Journal*, Vol 102, p446, (Abstract of a talk presented to the Christchurch Medical Society).

C.I. Watson, W.K Kennedy and R.H.T Bates (1990), "Towards A Computer-Based Speech Aid", In Proc. 3rd Australian International Conference on Speech Science and Technology", Melbourne, November, p234-239.

C.I. Watson, E. Terris, W.K. Kennedy and R.H.T Bates (1991), "Development and Evaluation Of A Computer-Based Speech Therapy Accessory", *Computer Technology For People With Special Needs*, Auckland, January.

C.I. Watson, W.K. Kennedy and R.H.T Bates (1991), "A Computer Based Speech Training Aid For The Speech Impaired: Development and Evaluation", In Proc. NELCON, (*New Zealand National Electronics Conference*), Palmerston North, August, p105-110.

C.I. Watson and J. H. Andreae (1992), "A Test To Assess The Remedial Worth Of A Computer-Based Speech Therapy Aid", In Proc. 4th Australian International Conference on Speech Science and Technology", Brisbane, December, p279-284.

CHAPTER 1

ASPECTS OF SPEECH

1.1 INTRODUCTION

The ability to speak to be understood is something most of us take for granted. Yet for those people with a speech disability it can be a very frustrating task. Fortunately through speech therapy it is possible to, if not cure, at least alleviate the speech impairment. The use of a visual feedback system, that can provide information on different aspects of speech, can be beneficial to the speech impaired in speech therapy.

This thesis is about the development and assessment of a computer-based speech therapy aid developed in the Department of Electronic and Electrical Engineering at the University of Canterbury. Chapter 1 of this thesis looks at speech, how it is produced, its segmental and suprasegmental units, how these units can be analysed acoustically and displayed on a visual medium and finally what are the disorders of speech. Chapter 2 presents a series of visual speech aids, it discusses the types of speech errors the aids are used for and how the information about the speech errors is displayed. Chapter 3 presents our aid, the Computer Aided Speech Therapy Tool (the CASTT) and describes each of the seven modules in the aid. Chapter 4 discusses the signal processing aspects of the CASTT, how the speech is processed and what analysis algorithms are used to calculate the speech features displayed in the CASTT.

In chapter 5 the evaluation of the CASTT by 15 speech therapists is discussed. It is shown how these evaluations led us to the Visual Display Test (VDT). The VDT is a test which establishes for which speech errors the CASTT can be expected to be an effective aid. The development and results of this test are given in chapter 6. The VDT and the therapists' evaluations suggested that two of the modules in the CASTT required new analysis algorithms. Chapter 7 investigates fricative recognition algorithms and their implications for a new module in the CASTT which would be used for fricative errors. Chapter 8 investigates a new vocal tract shape reconstruction algorithm; the algorithm is intended to replace an existing one in one of the modules of the CASTT. The final chapter concludes and discusses future work.

The purpose of this chapter is to provide the background to a speech aid which displays acoustic attributes of speech. In order to establish what is needed in a speech therapy aid it is necessary to discuss speech and language disorders and to outline speech and language therapy. People require speech and language therapy when they speak in a manner outside the accepted cultural norm and when their manner of speech or language is considered distracting or inappropriate. If people do not speak correctly it is due to improper use of the speech producing mechanisms, due to damage of the mechanisms or due to incorrect use of the language structure. The types of speech and language disorders will be discussed in section 1.5. The role of the speech language

therapist will also be discussed in this section.

However before speech and language disorders and speech-language therapy are discussed it is important to be aware of how speech is produced, what is the phonological structure of the New Zealand English language (the sound system) and what features of speech can be acoustically measured. These topics will be discussed in sections 1.2, 1.3 and 1.4 respectively.

1.2 ANATOMY AND PHYSIOLOGY OF SPEECH PRODUCING MECHANISMS

The human-being is able to produce a vast number of voluntary articulations which combine in various ways to form a complex oral communication system. Our ability to speak relies on our utilisation of existing anatomical structures, whose primary function is for the life supporting functions of eating or breathing. Speech production does have its own neural control system, however. To describe the process of speech production this section will look at the respiratory, laryngeal and supralaryngeal structures respectively and the corresponding actions of respiration, phonation and articulation. Finally it will outline, very briefly, the innervation systems that control and coordinate the vocal organs.

1.2.1 The Respiration Structure and its Function

The energy for speech is produced by the movement and the pressure of air in the lungs. The primary function of the lungs is to draw oxygen-rich air into the body so it can be absorbed into the blood and then to release the waste carbon dioxide from the body back into the air. The lungs are situated in the thoracic cavity. They are porous, spongy and elastic. There are no muscles in the lungs. The control of the air flow in both inspiration and expiration depends on the movements of the diaphragm and the muscles surrounding the ribcage, (see fig 1.1).

During inspiration of air the volume of the thoracic cavity increases by the diaphragm contracting and by the ribcage elevating and expanding. As the lungs increase in volume, the air at atmospheric pressure rushes in via the supralaryngeal and laryngeal structures (see fig 1.2). This action balances the drop in air pressure in the lungs due to expansion.

Expiration occurs when the air-pressure within the lungs exceeds that of atmospheric pressure. The air pressure within the lungs increases when the forces caused by the gravitational pull on the elevated ribcage, the elastic recoil of the lungs, the muscles around the ribcage pulling it down from its elevated position, and the compressed abdominal muscles pushing up on the diaphragm, are enough to cause the lung volume to reduce. When the pressure of the air in the lungs exceeds the atmospheric pressure the air is expelled. The air passes from the lungs, up the trachea to the larynx (see fig 1.1) and is expelled out of the mouth or nose.

In breathing the duration of inspiration and expiration are approximately equal. However in the production of English, expiration is sustained and controlled and inspiration is much shorter in duration than expiration. English speech is only produced in the expiration part of the respiration cycle. Speech produced in this manner is called egressive speech.

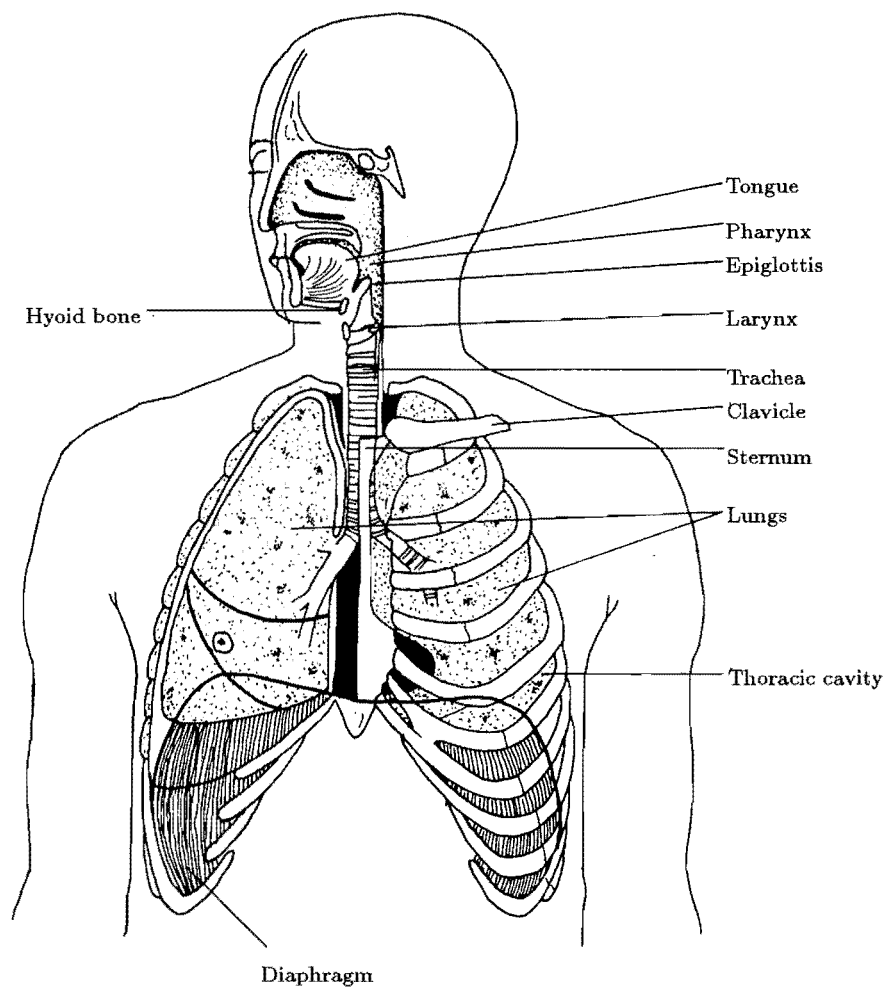


Figure 1.1. The Vocal Organs and associated structures (based on a figure in Crystal (1980) (p94)).

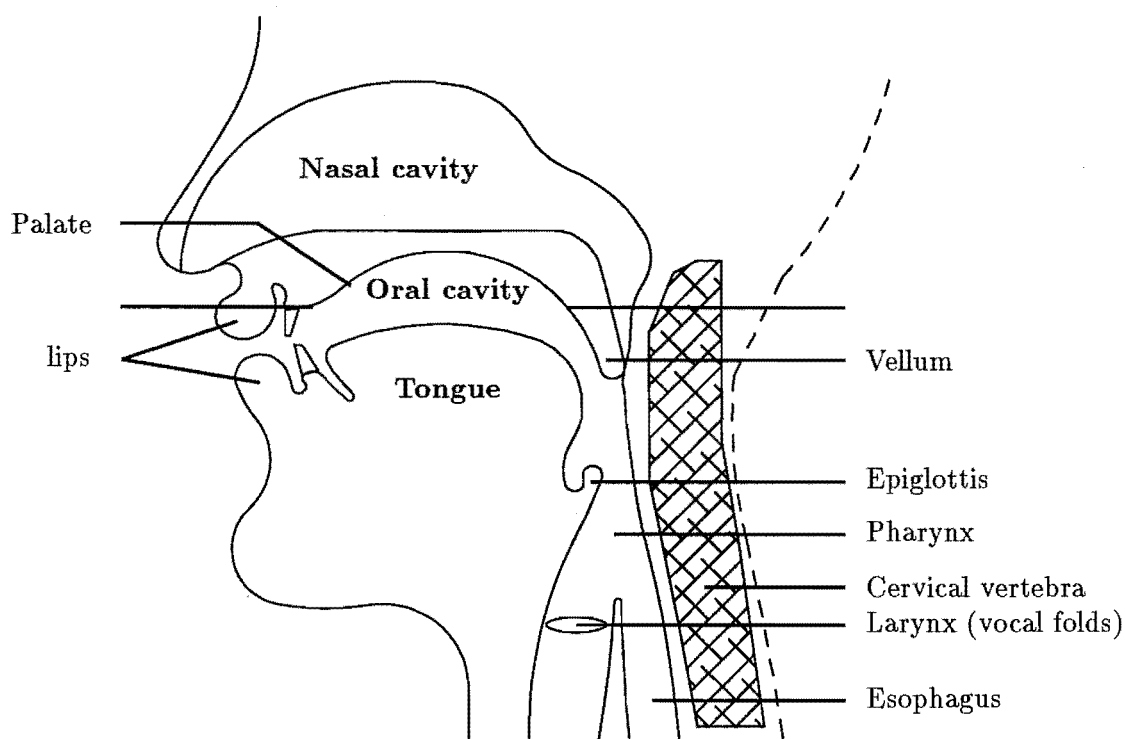


Figure 1.2. A cross-section of the head showing the larynx, oral and nasal cavity (courtesy of Andrew Elder).

1.2.2 The Laryngeal Structure and Phonation

The larynx is the organ of phonation. Its primary function is to protect the lungs and trachea from solid and liquid materials taken into the mouth. The protection occurs due to the sealing actions of the epiglottis and the vocal folds (see fig 1.2).

Figure 1.3 is the view from (a) the side of and (b) above the larynx. The larynx consists of 9 cartillages; the largest of these is the thyroid cartilage, which is attached to and pivotted from the cricoid cartilage which in turn is attached to the trachea. At the back of the cricoid cartilage are the two arytenoid cartillages. Extending from each of the arytenoids to the inside of the thyroid are the vocal folds. The folds are mainly muscle tissue except right at the very edge where they are ligament (Levelt, 1989). The movement of the vocal folds is caused by the many muscles in the larynx and by the movement of the arytenoids. The gap between the vocal folds is referred to as the glottis. The tension of the vocal folds can also be changed by the muscles in the larynx. The vocal folds are also known as the vocal cords.

During normal breathing the vocal folds are held apart, as seen in fig 1.4(a). In phonation the vocal folds are brought together (see fig 1.4(b)). Phonation occurs on expiration. As the air passes through the larynx the vocal folds are set into vibration and thus act as the excitation source of speech. In English the main sound source is the vibrating vocal folds. The accepted theory of phonation is the aerodynamic-myoelectric theory (Hardcastle, 1976, p83). It states that the valve-like vocal folds seal off the expiratory air flow. Consequently the subglottal air pressure builds up until the elastic

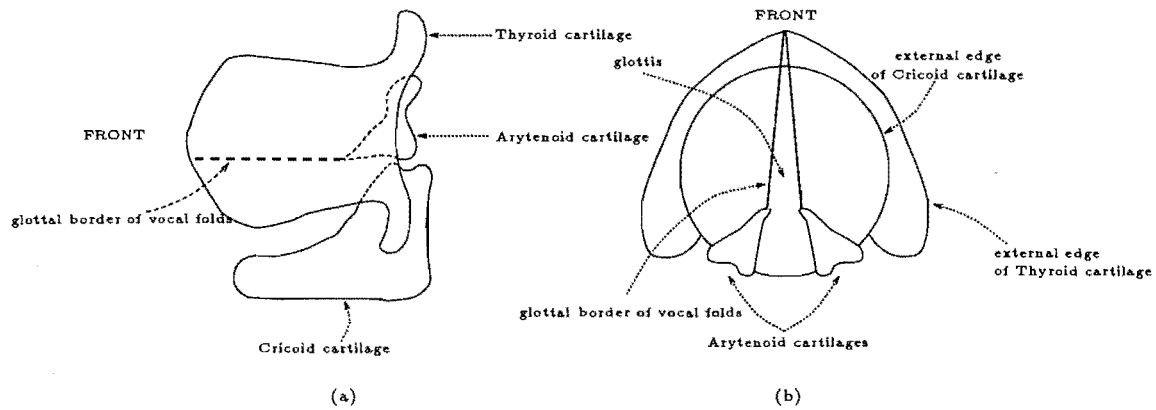


Figure 1.3. The view from (a) the side and (b) from above of the larynx (based Laver (1980) (p102)).

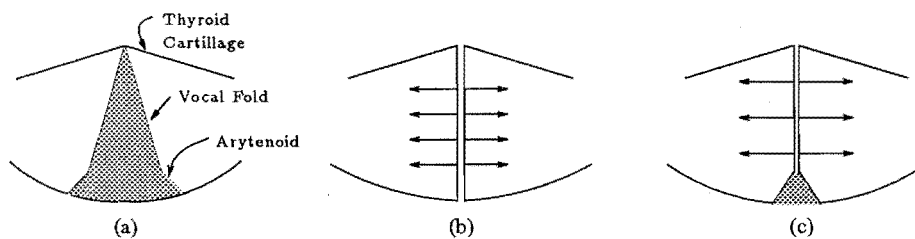


Figure 1.4. A highly stylised drawing of the glottis in three modes (a) inspiration (b) normal phonation and (c) breathy voice (based on Clark and Yallop (1990) (p42)).

vocal folds are pushed apart, forming a narrow opening. As the air flows through the narrow opening (the glottis) the velocity of the air molecules increases, so the glottal pressure decreases. The subsequent loss of pressure enables the folds to return to their unstressed state and close again. Once the folds are closed the expiratory air flow is stopped and the cycle begins again.

The frequency of the vibration of the larynx in phonation is the fundamental frequency of the speaker. The perceptual correlate of this frequency is pitch (Clark and Yallop, 1990). The fundamental frequency of the vocal fold vibration varies during speech. Thus the pitch of the voice never remains the same for an appreciable time (Fry, 1979). The variations in pitch provide the intonation patterns of speech (Clark and Yallop, 1990). The frequency of the vibration of the vocal folds is dependent on the subglottal air pressure and on the length, tension and mass of the vocal folds. There is a wide variability in the fundamental frequency range but the general range for speakers of English is : 80 - 130 Hz for adult men; 150 - 300 Hz for adult women and 200 - 500 Hz for children (Clark and Yallop, 1990).

There are three auditory parameters associated with phonation (and indeed with speech as a whole): pitch, loudness and quality. They are all inter-related (Clark and Yallop, 1990). Pitch has been discussed above. Loudness is, for the most part, a perceptual response to the amplitude of the speech signal (Lieberman and Blumstein, 1988). The intensity of the acoustic excitation produced by phonation is controlled by the air pressure. This is kept fairly constant throughout phonation. The muscles in the respiratory system are used to maintain the constant air pressure. There is, however, sufficient variation in the air pressure to allow for variations in intensity (Clark and

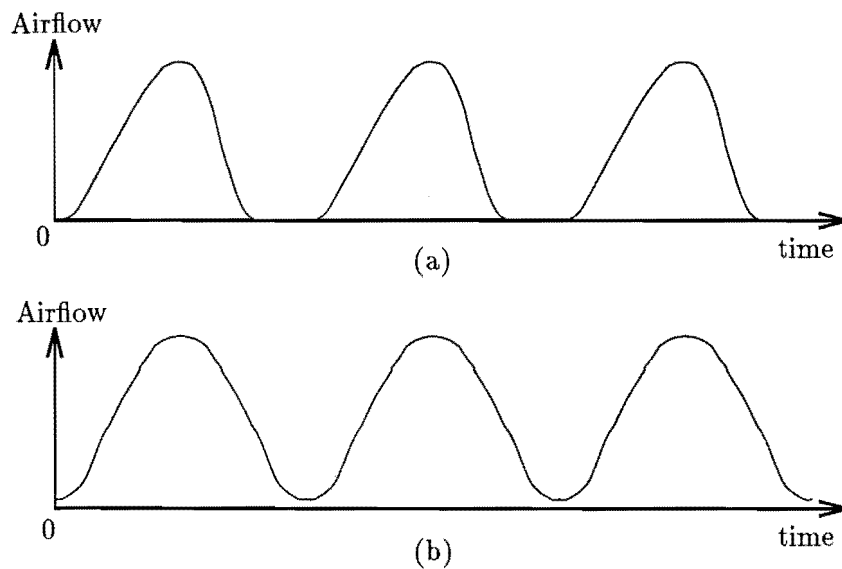


Figure 1.5. Stylised airflow waveforms in the glottis produced in (a) normal phonation and (b) breathy phonation.

Yallop, 1990).

Quality, which is also known as timbre, is the perceptual correlate of the spectral content of a sound. The spectral content is related to the sound source and the acoustic chamber from which the sound was produced. The quality of speech will change if there is a change in the vocal tract shape or if there is a change in the sound source. To distinguish between the two different types of quality in this thesis, the quality of speech associated with the sound source will be referred to as the timbre of speech.

The timbre of speech can be changed by altering the vocal folds. The vibration of the vocal folds provides the sound source and the vocal tract shape provides a resonant chamber. Figure 1.5 (a) is a stylised glottal waveform produced in normal phonation. It is this waveform which is resonated by the variable vocal tract to produce speech. The tension of vocal folds can be altered, through the muscular action of pivoting the thyroid cartilage forward on the cricoid, for example, or through the arytenoid cartilages pivoting at a different angle than in normal phonation. By altering the vocal folds, the glottal air flow waveform changes. For example, fig 1.4 (c) gives the vocal folds' mode in breathy speech. Whilst the folds can still vibrate, a portion of the glottis is always left open. Figure 1.5 (b) is a stylised glottal waveform produced by breathy speech. It can be seen that it looks slightly different from the glottal wave form for normal speech. The spectral content of these two waves will be slightly different. Thus when glottal waveforms of normal phonation and breathy phonation are resonated in the same vocal tract the resulting speech waveform will be different.

1.2.3 The Supralaryngeal Structures and Articulation

The supralaryngeal structures (the vocal tract), (see figure 1.2), filter the acoustic energy generated at the larynx or at a point of constriction to give the speech its quality. As we talk we continually change the shape of the vocal tract, thus producing the different sounds that constitute speech. The vocal tract comprises the pharyngeal,

nasal and oral cavities.

The shape of the vocal tract can be altered by changing the position of the tongue, mandible, lips, velum and pharynx (see fig 1.1). The tongue has the biggest effect on changing the vocal tract shape. It is mainly muscle tissue with a mucous membrane cover and is attached to the skull, mandible and hyoid bone (see fig 1.1). Two thirds of the tongue is in the oral cavity and the rest is in the pharyngeal cavity.

The primary use of the tongue is to position food in the mouth for chewing and to transport the masticated food into the throat for swallowing. It is very flexible. It can be protruded out of the mouth, lifted or depressed in the oral cavity and the tongue tip can be moved up and down. The tongue can change its shape and become long, narrow, wide, flat, concave or convex. All these movements are utilised in articulation (see section 1.3).

The mandible is attached to the hyoid bone and tongue by means of muscle. Its life-supporting function is to chew and grind food. It is able to move sideways, backward, forward, up and down. These last two movements are used in speech production (Levelt, 1989, p430). The mandible can alter the position of the tongue and the lower lips. These actions change the volume of the vocal tract and hence its resonance.

The lips comprise mainly tissue, blood vessels, glands, nerves and muscle. They encircle the mouth. Their primary function is for sucking and eating. The facial muscles around the lips enable pursing, rounding, elongation and closing of the lips. Lip sounds are one of the first sounds babies make (Skinner and Shelton, 1978, p65).

The velum, also known as the soft palate is soft and flexible. It can be thought of as a flexible extension to the hard palate. It has two life-supporting functions. In breathing, it is lowered to allow air flow through the nasal cavity. In swallowing the velum rises, closing the velopharyngeal passage to seal off the nasal cavity. When the velum is lowered during speech the nasal cavity becomes a resonator. Most sounds in New Zealand English are not nasalized and the velopharyngeal passage is normally closed during speech.

The final vocal organ that can change the shape of the vocal tract is the pharynx, which is a flexible muscular tube. Its first primary function is pushing the masticated food from the mouth into the esophagus. Its second is providing a passage from the nose and mouth for the air to flow to and from the lungs.

1.2.4 The Innervation System and Speech Control

One of the most important requisites for intelligible speech is the simultaneous control and coordination of the respiratory, laryngeal and supralaryngeal structures. Speech production has its own neural control, which in turn is controlled and coordinated by the innervation system. The nature and presence of the respiratory, laryngeal and supralaryngeal structures in human-beings alone is not enough to enable the voluntary articulatory manoeuvres that constitute speech. It would be impossible to talk without the specialized human-specific brain mechanisms that synergise the respiratory, phonatory and articulatory systems (Lieberman and Blumstein, 1988). The coordination and control of the speech mechanism is a complex task. In order to articulate there are 15 supralaryngeal structures, over 100 muscles and many bodily systems (such as skeletal, muscular, skin, digestive, respiratory etc) to be controlled.

The neural control of speech involves both the Central Nervous System and the

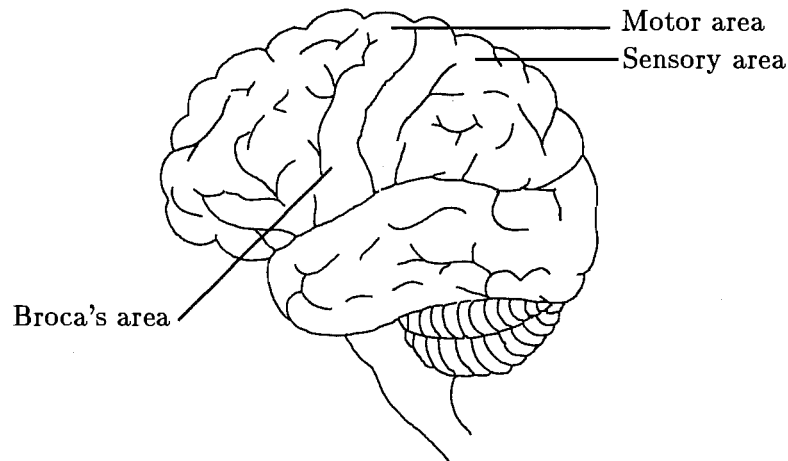


Figure 1.6. The view of cortex showing areas of brain used in speech production (figure courtesy of Alison Dingle).

Peripheral Nervous System. The Central Nervous System comprises the spinal cord and the brain. The Peripheral Nervous System, which connects various parts of the body to the brain and spinal cord, consists of cranial and spinal nerves. Figure 1.6 shows the parts of the cerebral cortex used in speech production. The motor area of the cortex controls bodily movements including those used in phonation and articulation. The sensory area receives information from the muscles involved in the bodily movements, the information being used to guide production of normal speech. The Broca area situated in the cortex of the left hemisphere enables voluntary articulation (see fig 1.6). It controls the motor and sensory areas for speech production and is essential for speech production.

The process of how speech is planned in the brain before articulation, eg. in allophonic units or syllable units etc, is still conjecture. There are many theories on this topic but they are beyond the scope of this thesis. The reader is referred to Levelt (1989) for more information on this topic.

1.3 THE NATURE OF SPEECH

We know instantly when someone speaks in a manner which is outside our social norm. Errors in the features of their speech enable us to make this judgement. The purpose of this section is to discuss these features. However it must be stressed that this in no way implies that natural speech can somehow be attained just by concatenating and superimposing these features together.

The first deconstruction of speech will be to describe it in terms of its segmental features and its suprasegmental features.

1.3.1 Segmental Aspects Of Speech

Speech sounds can be described phonologically, mechanically (how they are produced) and acoustically. In this section we will be describing them phonologically and mechan-

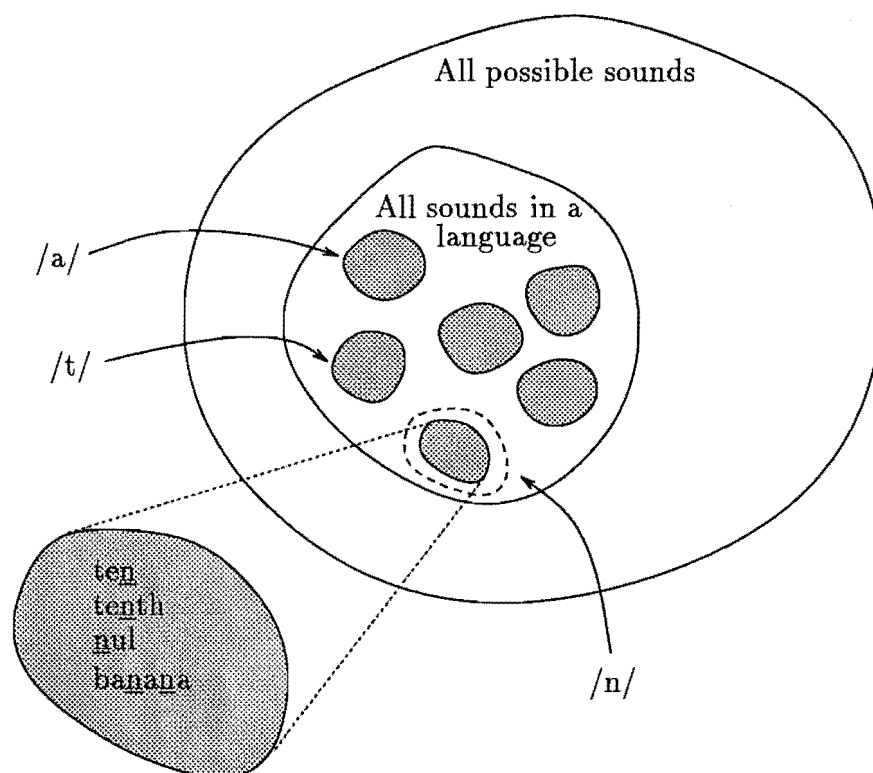


Figure 1.7. A graphical representation of what constitutes a language.

ically. It is difficult to describe them in these ways without referring to their acoustic description but this is the approach that will be taken.

1.3.1.1 Phonological Structure

A phonological approach to language “requires that we examine the linguistic value of the sounds and that we list the phonemes - the system of sounds considered as elements which serve to distinguish the meaning of words” (Jakobson, 1990). Ladefoged (1975) went on to say that phonemes are “the abstract units that form the basis for writing down a language systematically and unambiguously”. Closely associated with phonology is phonetics. A phonetic approach to language aims at “drawing up an inventory of the sounds of a language considered simply as motor and acoustic phenomena” (Jakobson, 1990).

The above concepts are illustrated in fig 1.7. Imagine that phonetic space is the set of all the possible sounds that can be produced by the human speech mechanisms. The sounds of a particular language are a much smaller subset in the phonetic space. The sounds of a language can be grouped into even smaller regions which lie in the phonetic space of the language. Each of these regions represents the phonemes of the language. All the sounds which lie in a phoneme region are allophones of it. The occurrence of an allophone in spoken language depends on the sounds it precedes and follows and the accent/dialect of the speaker. A few words, in which some of the different allophones of the English phoneme /n/ occur, are listed in figure 1.7.

There are 43 phonemes in New Zealand English. Table 1.1 lists the symbols for

| The Phonemes of New Zealand English | | | |
|-------------------------------------|------------------|------|----------------|
| /p/ | <u>p</u> at | /i/ | pe <u>a</u> t |
| /t/ | <u>t</u> art | /u/ | bo <u>o</u> t |
| /k/ | <u>c</u> ard | /I/ | p <u>i</u> t |
| /b/ | <u>b</u> at | /ʊ/ | pu <u>u</u> t |
| /d/ | <u>d</u> art | /e/ | pe <u>e</u> t |
| /g/ | <u>g</u> uard | /ɜ/ | pe <u>r</u> t |
| /f/ | <u>f</u> an | /æ/ | ca <u>a</u> t |
| /θ/ | <u>th</u> in | /ʌ/ | bu <u>u</u> t |
| /s/ | <u>s</u> ea | /ɔ/ | po <u>o</u> t |
| /ʃ/ | <u>sh</u> in | /a/ | pa <u>a</u> t |
| /h/ | <u>h</u> ea | /ɒ/ | po <u>o</u> t |
| /v/ | <u>v</u> an | /ei/ | da <u>y</u> |
| /ð/ | <u>th</u> en | /iə/ | he <u>r</u> e |
| /z/ | <u>z</u> ea | /ai/ | di <u>e</u> |
| /ʒ/ | lei <u>z</u> ure | /eə/ | the <u>r</u> e |
| /tʃ/ | <u>ch</u> in | /ɔi/ | bo <u>y</u> |
| /dʒ/ | gi <u>z</u> | /ʊə/ | to <u>u</u> r |
| /m/ | <u>m</u> ad | /ɔʊ/ | go <u>o</u> |
| /n/ | <u>n</u> od | /aʊ/ | co <u>o</u> |
| /ŋ/ | si <u>ng</u> | | |
| /l/ | <u>l</u> ad | | |
| /r/ | <u>r</u> od | | |
| /w/ | <u>w</u> ad | | |
| /j/ | <u>y</u> ard | | |

Table 1.1. The 43 phonemes in New Zealand English and an allophonic example of each phoneme in a word (this table is from a handout for a 200 level linguistics course given by the Department of Linguistics at the University of Canterbury).

transcribing phonemes and it also gives an allophonic example of each phoneme in a word.

When a language is broken down into a series of sounds through phonetic analysis the sounds are represented between square brackets, []. When a language is broken down into the smallest linguistic units (phonemes) which can distinguish the meaning of words, through phonological analysis, the units are represented between slashes, / /. Thus the phoneme /n/ may be represented by the alveolar [n] (as in “ten”) or the dental [ɲ] (as in “tenth”). Phonetic analysis would class [n] and [ɲ] as two quite different sounds. It is only because they are part of the same phonological set that the “n” symbol appears in both. Thus there is an implicit suggestion of meaning when one does a phonetic transcription of an unknown language (based on the language of the transcriber).

If two languages are compared from an acoustic and motor point of view their sounds could be identical. However the way in which they are grouped into phonemes could be quite different. Two sounds which are allophones in one language need not be in another. For example, the sounds [l] and [r] are phonemes in all forms of the English

language (ray-lay, furry-fully¹). In Korean they are allophones of a single phoneme. In Korean the phoneme represented by [l] appears at the beginning of words and the phoneme represented by [ɾ] occurs at the end of syllables. Thus for a Korean learning English, “rule” would be confused with “lure” (Jakobson, 1990).

We have now established the notion of the phoneme as a linguistic unit which distinguishes meaning. Next we must consider the sounds which lie within the phoneme regions. We have to establish whether, in the production of any allophone of a phoneme, there is something inherent in its production which distinguishes it from the production of an allophone of another phoneme. That is, is there something inherently “n-ish” about the production of an allophone of /n/ which distinguishes it from the production of an allophone of /a/ or /t/ say.

1.3.1.2 Articulation

The groups of allophones which lie in each of the phoneme regions can be described in terms of articulation. The sounds of New Zealand English, or indeed any form of English, are divided into two categories, consonants and vowels. This section will look at these sounds.

Consonants are speech sounds formed by the bringing together of articulators creating an obstruction (either partial or total) of the air flow from the lungs. Consonants can be described in terms of the place and manner of this obstruction (Ladefoged, 1975). The part of the vocal tract which is obstructed is the “place of articulation”. In English there are nine places of articulation according to Hawkins (1984). These are: Bilabial (constriction between the upper and lower lips); Labio-dental (constriction between the bottom lips and the edges of the incisors); Dental (constriction between the incisors and the tip of the tongue); Alveolar (constriction between the alveolar ridge and the tip of the tongue); Post-alveolar (constriction between the tongue blade and the post alveolar region); Palato-alveolar (constriction between the tongue blade and the region between hard palate and the alveolar ridge); Palatal (constriction between the blade of the tongue and the hard palate); Velar (constriction between the back of the tongue and the soft palate) and glottal (constriction of vocal folds) (Hawkins, 1984), (Clark and Yallop, 1990).

The manner in which the blockage of air is caused is called the “manner of articulation”. It, like the place of articulation, also influences the production of a sound. The manners of articulation in English are plosives, fricatives, affricatives, nasals, laterals, frictionless continuants and glides. To produce plosives the vocal tract is completely closed for a moment and then it is abruptly opened. The closing of the vocal tract causes a build up of air pressure. On opening the lips the air rushes out causing a short burst of sound. An example of an English plosive is [p]. Fricatives (e.g. [s]) are caused by partial closure of the vocal tract at some point causing turbulence in the air flow. Affricatives (e.g. [tʃ]) are caused by the combination of a plosive followed by a fricative. The build up of air is released more slowly in an affricative than in a plosive.

Plosives, fricatives and affricatives can be voiced or unvoiced. When these sounds are voiced there are two sound sources, one at the glottis and one at the constriction.

¹In the original example Jakobson used fur-full, however this example does not hold for New Zealand English. Whilst in American English there is a final [ɾ] in pronunciation of “fur”, there is no final [ɾ] for New Zealand English pronunciation: the word “fur” is pronounced [fɜː].

| | bilabial | labiodental | interdental | alveolar | post-alveolar | palto -alveolar | palatal | velar | glottal |
|------------------------------|--------------------------------|--------------------------------|----------------------------------|----------------------------------|---------------|-------------------------------------|----------------|-----------------------------------|----------------|
| Plosive | [p] as in pat [b] as in bat | | | [t] as in tart [d] as in dart | | | | [k] as in card [g] as in guard | |
| Fricative | | [f] as in fan [v] as in van | [θ] as in thin [ð] as in then | [s] as in seal [z] as in zeal | | [ʃ] as in shin [ʒ] as in leisure | | | [h] as in heel |
| Affricative | | | | | | [tʃ] as in chin [dʒ] as in gin | | | |
| Nasal | [m] as in mad | | | [n] as in nod | | | | [ŋ] as in sing | |
| Laterals | | | | [l] as in lad | | | | | |
| Fricationless Continuents | | | | | [r] as in rod | | | | |
| Semi-Vowel | [w] as in wad | | | | | | [j] as in yard | | |

Table 1.2. Examples of allophones of New Zealand English described in terms of place and manner of articulation.

The sound source due to the glottis is the periodic vibration of the vocal cords. The sound source at the point of constriction is due to the turbulence in the air flow caused by the constriction. When a sound is unvoiced the air turbulence caused by the constriction is its only sound source, the glottis being kept open. When two plosives, fricatives or affricatives are described as a voiced/unvoiced pair, it means they are sounds that have the same place and manner of articulation but differ only in the vibration or nonvibration of the vocal cords. The timing of the start of phonation relative to the supralaryngeal behaviour is called the voice onset time (Clark and Yallop, 1990). Voice onset time is an important cue in the distinction of voiced/unvoiced pairs in English (Fry, 1979, p136).

Nasal, lateral, frictionless continuants and semi-vowels all have just one sound source - the periodic glottis. In nasals sounds (e.g [m]) the velum is lowered and the nasal and oral cavities are coupled and the exit to the oral cavity is completely sealed off. Both cavities are used as resonant chambers. Laterals are sounds caused by air passing either side of the elevated tongue tip. In English there is one lateral, [l]. In frictionless continuant sounds the tongue is curled upwards and back. The middle of the tongue is concave. There is one frictionless continuant in English, namely [r]. The final manner of articulation is the semi-vowel, which is also known as glides. The semi-vowel is made by two articulators being brought very close together causing some perturbation in the air flow. No actual contact is made between the articulators. Semi-vowels are of very short duration and cannot be sustained. In English the semi-vowels are [w, j]. The above explanations of all the manners of articulation were based on Hawkins (1984) .

Table 1.2 classifies an allophone of each of the New Zealand consonant phonemes in terms of manner and place of articulation. An example of the allophone in a word is also given.

The places and manners of articulation are not limited to those mentioned above. There are many other different ways sounds can be produced. However they are not used contrastively in the production of New Zealand English. The speech analysed and discussed in this thesis is English or more specifically New Zealand English and its perturbations. Thus, other places and manners of articulation which occur in other languages are ignored.

In the production of vowels the air flow is virtually unobstructed, no articulators forming points of constriction. Vowels are vocalic sounds, that is they are sounds created by egressive pulmonic air flow (ie. from the lungs) causing the vocal cords to vibrate. The sound generated in the larynx is modified by the shape of the vocal tract.

| New Zealand Vowels | | | |
|--------------------|-----------------|----------------|----------------|
| Jaw Position | Tongue Position | | |
| | FRONT | CENTRE | BACK |
| Close | [i] as in peat | [u] as in boot | [ɔ] as in port |
| | | [ɜ] as in pert | |
| | [e] as in pet | [ɪ] as in pit | [ʊ] as in put |
| open | [æ] as in cat | [ʌ] as in but | [ɒ] as in pot |
| | | [a] as in part | |
| | | | |

Table 1.3. The New Zealand Vowel sounds, with examples of the sounds in words.

| New Zealand Diphthongs | |
|------------------------|------------------|
| Closing | Centering |
| [ei] as in day | [iə] as in here |
| [ai] as in die | [eə] as in there |
| [ɔi] as in boy | [ʊə] as in tour |
| [ɔʊ] as in go | |
| [aʊ] as in cow | |

Table 1.4. An allophonic example of each New Zealand Diphthong with examples of the sound in a word.

The two most important articulators in vowel production are the tongue and the lips. There are two types of vowels in English, monophthongs and diphthongs.

Traditionally monophthong vowels are described in terms of tongue height and position, which in turn are traditionally referenced by the highest point on the dorsum of the tongue. Lip configuration is also important (Skinner and Shelton, 1978, p67). Allophonic examples of the vowel phonemes for New Zealand English, presented in terms of tongue height and position, are given in Table 1.3.

Diphthong vowel sounds occupy a single syllable but the articulation involves tongue movement from the position of one vowel to the position of another. They are described in terms of the movement of the tongue hump. Diphthongs in which the tongue hump moves towards the extremes of the oral chamber (e.g top front) are called closing diphthongs. Diphthongs in which the tongue hump moves towards the centre of the oral chamber are called centring diphthongs. Table 1.4 gives an allophonic example for each of eight New Zealand English diphthongs.

We have just discussed speech in terms of segmental features (consonants and vowels). However, it is important to remember that speech does not consist of a series of target articulations, linked together by movement of the articulators from one target to the next. There is considerable overlap in the articulation of the sounds. The articulations of a sound are affected by the sounds which precede and follow it. Thus while there may be an ideal way of articulating a sound it is subject to modification in continuous speech (Clark and Yallop, 1990).

The overlapping of adjacent articulations is called co-articulation (Ladefoged, 1975). An example used by Ladefoged (1975) of how co-articulation can affect the place of articulation is the [k] sound in “key” and “caw”. The articulation of /k/ is palatised in “key”, the tongue hump being much closer to the front of the mouth than in the

realisation of /k/ in “caw” as a post-velar. English consonants often vary in their place of articulation and become more like the sound which follows them (Ladefoged, 1975).

1.3.2 Suprasegmental Aspects Of Speech

In the last section we discussed the segmentals of speech, the sounds produced in a language. There are also suprasegmentals of speech, which have been described as “the melody of speech” (Nation and Aram, 1977). There are many features which contribute to the “melody of speech” such as variations in loudness, pitch, rhythm, fluency, duration, speaking rate and phrasing. On a linguistic level the most important suprasegmental features are intonation, stress and duration.

Intonation is the name given to the variations of pitch in an utterance (Hawkins, 1984). It is the linguistically significant function of pitch at the sentence level. Intonation does not change the meaning of the lexical items in speech but it forms part of the meaning of the entire utterance (Lehiste, 1976). Thus intonation cues are used to distinguish between the question “you are wearing wool?” (the pitch rises at the end of this sentence) and the statement “you are wearing wool!” (the pitch falls at the end of this sentence).

Stress functions linguistically at the word and sentence level. Pitch, loudness and duration of sounds (realized in duration of syllables) are all components of stress. In English a more stressed syllable is longer in duration than a less stressed syllable. In the perception of stress, for English, the most important feature is pitch prominence, followed by duration, then loudness (Lehiste, 1976).

At the word level stress can distinguish between many verb/noun combinations in English (for example compare “to *contract*” with “a ” *contract* or “to *permit*” with “a *permit*”). The verb is stressed on the second syllable and the noun is stressed on the first. At the sentence level, stress does not change the meaning of any lexical item but it increases the relative prominence of one of the words (Lehiste, 1976). For example, in the phrase “he was a French teacher” if “French” is emphasized it implies the teacher teaches French, while if “teacher” is emphasized it implies the teacher is from France.

The suprasegmental features of speech also have nonlinguistic and paralinguistic functions. It is important that these features, as well as the linguistic features, are appropriate for speech to sound “normal”. The nonlinguistic suprasegmental features of speech are those of voice quality and timbre. These features reflect the nature of the larynx and the vocal tract; for example the timbre of phonation is determined by the anatomy and tension of the larynx. The nonlinguistic features of speech do not form part of the functional system of language. Paralinguistic features of speech are those that provide information about the speaker’s emotional state (anger, fear, excitement) and the speaker’s attitude to the listener and the message (submissiveness, nervousness, authoritativeness).

1.4 THE ACOUSTIC FEATURES OF SPEECH

This section will show that acoustic representations of loudness, pitch and quality of speech can be displayed visually. The section is intended only to be an introduction to the acoustic features of speech. These features and how to calculate them will be discussed in more detail in later chapters where it is appropriate to present them.

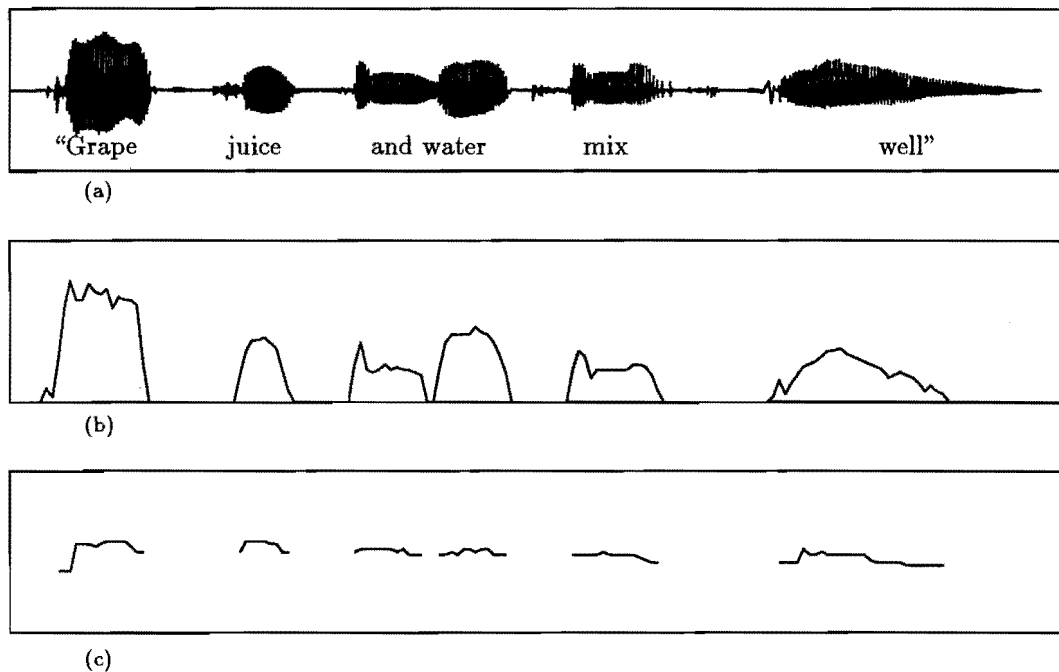


Figure 1.8. The (a) time domain wave form of the utterance "Grape juice and water mix well" (b) the intensity variations of the utterances and (c) the fundamental frequency variations.

The features of speech so far discussed in this chapter are those pertaining to language. We have seen, for instance, how sounds combine to convey meaning, how the loudness and the pitch of a voice combine to give timbre and also to provide meaning. We will now shift our emphasis from looking at the linguistic function of speech to looking at its acoustic features. Acoustically speech is a pressure wave which propagates longitudinally through the air. The pressure wave is caused by the expired air flow from the lungs through the larynx and/or the articulators in the vocal tract.

The three factors which provide the easiest method of differentiating speech sounds are the perceptual features of loudness, pitch and quality (Ladefoged, 1972). These concepts have already been introduced in sec 1.2.2. In this section we will look at the acoustic realization of these auditory parameters.

Loudness is related to the subglottal air pressure. It is related to the amplitude of the time domain waveform, but it is more correct to say that it is related to the energy of the sound (also known as its intensity) (Fry, 1979). The greater the amplitude of the time domain waveform, the greater the energy of the sound and the louder the sound appears. Pitch is the perceptual correlate of the fundamental frequency of the vocal fold vibration of the speaker. Figure 1.8(a) gives the time domain waveform of the phrase "Grape juice and water mix well". Figure 1.8(b) gives the intensity variations of the phrase and Figure 1.8(c) gives the fundamental frequency variations.

The quality of a sound is the perceptual correlate of its spectral content. The spectral content of a sound is affected by how the position of the articulators alters the waveform of the sound source (the larynx vibration and/or air turbulence caused by a constriction in the vocal tract). The combination of the sound source and the articulator

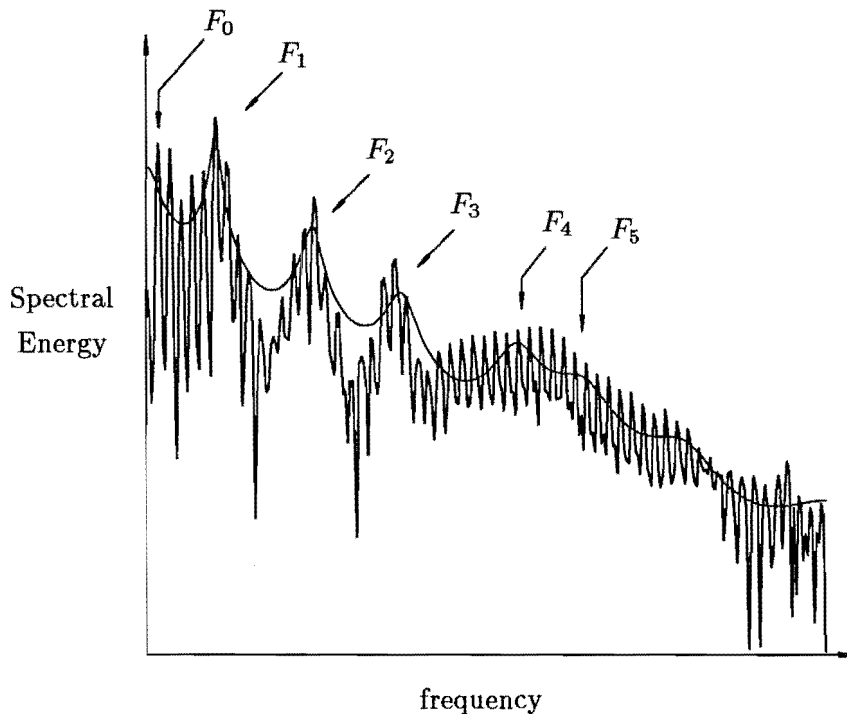


Figure 1.9. An example of an amplitude spectrum of speech, the peaks are the formants (this figure is courtesy of Andrew Elder).

positions leads to the source filter model of speech production, which is one of the most commonly used models for speech production. It originates from the mid-nineteenth century and still guides research on speech today (Lieberman and Blumstein, 1988, p1).

In the source-filter model for sounds in which the vibrating larynx is the only sound source (e.g. vowels and semi-vowels), the glottal waveform is assumed to be independent of the sound being produced. As the glottal pulse propagates down the vocal tract, certain frequency components of the glottal wave resonate. These components are known as the formants. The formants of a sound are the positions in its spectrum where it has the greatest acoustic energy and they can be identified in the spectrum from the positions of the energy peaks, as indicated in fig 1.9. The shape of the vocal tract determines which frequency components resonate. The short hand for the first formant is F_1 , for the second is F_2 etc. The fundamental frequency is indicated by F_0 .

Sounds for which at least one of the sound sources is due to air turbulence caused by a constriction, do not have formants in their spectrum.

A spectrogram of a speech signal shows how the spectrum of speech changes over time. For example fig 1.10 shows the spectrogram of the utterance “berries”. It is calculated from sequential segments of speech. The horizontal axis shows time and the vertical axis shows frequency. The colour scale (the grey scale) shows the energy of the frequency components. The darker the colour the higher the energy of the component. The three formants of the sound [e] in “berries” can be seen clearly. The positions

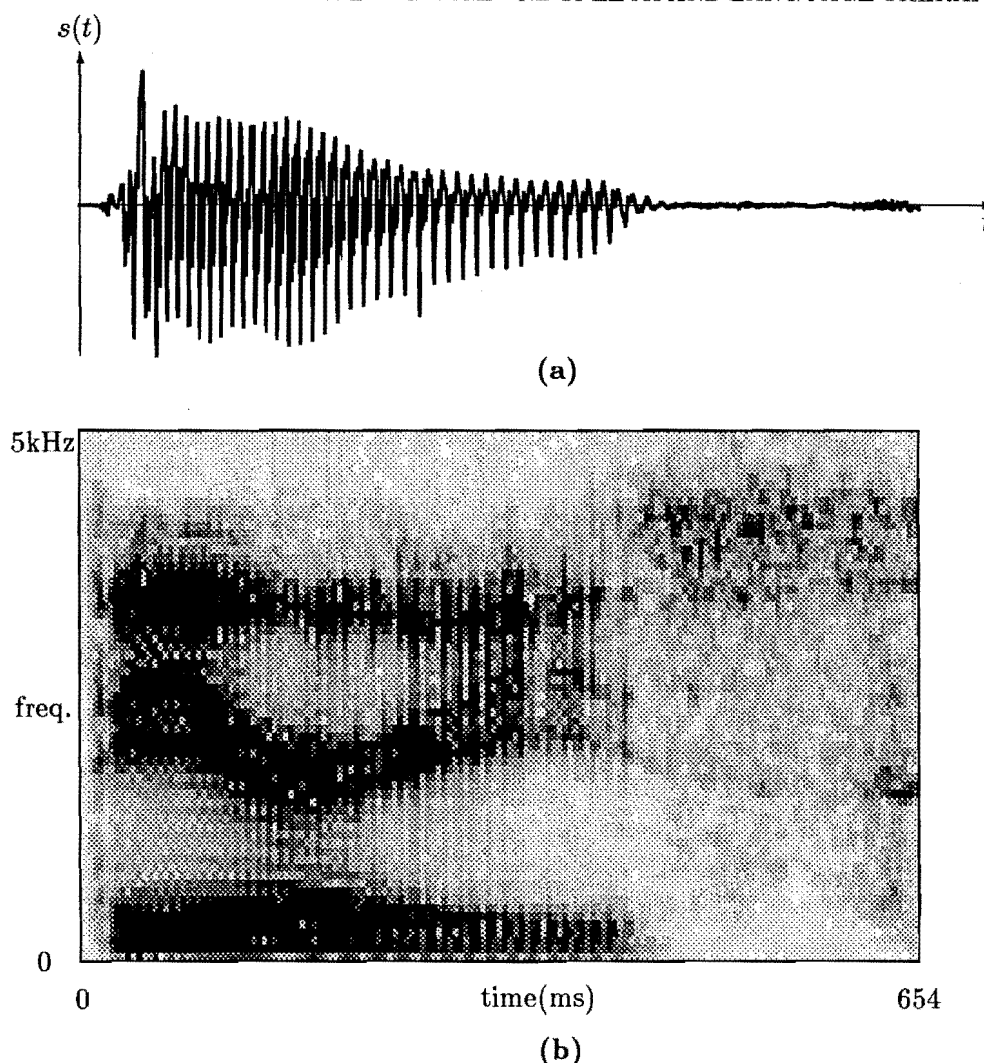


Figure 1.10. The time domain wave form of the utterances “berries” and the spectrogram of the utterance (this figure is courtesy of Andrew Elder).

of the formants change due to the co-articulation because the vowel follows [b] and precedes [r]. The formant shift can also be seen as the utterance shifts through the [r] sound to the [i] sound. It can also be seen that the spectrogram portion of the [z] in “berries” is unstructured. In addition it is the higher frequency components of [z] that have the most energy, unlike the non fricative sounds.

1.5 DISORDERS OF THE SPOKEN LANGUAGE AND SPEECH AND LANGUAGE THERAPY

We know immediately if someone speaks in a manner outside the accepted social norm. What is it about the way they speak which enables us to make that judgement? We are able to conclude from sec 1.3 that the differences would be manifested in the articulations, the phonological structure and the suprasegmental features of the speaker. These differences will be referred to as “disorders” because we are primarily interested in speakers having disabilities in the production of speech.

The disorders of spoken language fall into two groups; central pathologies and production pathologies (the nomenclature and division are those of Crystal (1980)). The central pathologies are those disorders of spoken language which originate in the cen-

tral nervous system (Crystal, 1980). Examples of these disorders are Aphasia, Auditory Agnosia, Dyspraxia and Apraxia. Aphasia is “a communication disorder caused by brain damage and is characterised by complete or partial impairment of language comprehension, formulation or use” (Crystal, 1980). Auditory Agnosia is “the inability to recognize and differentiate sounds, words etc” (Crystal, 1980). Finally Dyspraxia and Apraxia are “disruptions in the ability to produce purposeful motor responses” (Crystal, 1980).

The pathologies of production can be broken into three groups; the disorders of fluency, voice disorders and the disorders of articulation. Disorders of fluency cause specific disturbances in the rhythm and timing of speech. The most common of these disorders are stuttering and cluttering.

Voice disorders are manifested in the loudness, pitch and timbre of the speaker. They can cause the pitch range to be excessively high or low, or excessively wide or narrow. There might be a high level vibrato in the pitch or abnormal use of the pitch register. The loudness of the speech can be excessively loud or weak or be controlled erratically. The timbre of the voice can become breathy, abnormally nasal (hypernasal) or denasalised (hyponasal). In addition, when a source other than the larynx is used for phonation the timbre may sound abnormal.

The final set of disorders are those of articulation. Crystal (1980) described articulation disorders as “interference with the phonetic realization of the abstract units that constitute the linguistic system”. The errors of articulation are the addition of sounds (e.g. “bag pronounced as [bwag]), the omission of sounds (e.g. “balloon” pronounced as [bun]), the substitution of sounds (e.g. “bin” pronounced as [pin]) and the distortion of sounds (e.g. when the [s] sound is distorted due to a lisp) (Crystal, 1980) (Skinner and Shelton, 1978).

It is possible for people to have disorders which affect the phonological system of a language. They are able to produce all the sounds correctly but they have difficulty in organizing the sounds in order to make meaningful contrasts. It is also possible for people to have disorders which affect the phonetic structure of speech only. The phonological structure of a language remains intact but due to some damage to the speech-producing mechanisms the allophonic variations of the phonemes are misarticulated. Many disorders are a combination of the two.

Those people with a spoken language disorder are often sent to speech and language therapists. The speech and language therapists are specially trained to help the speech impaired person talk in a manner as closely as possible to the accepted cultural norm. A therapist's clients may be adults or children. The therapist evaluates the client's speech, diagnoses the disorders, plans a treatment to remediate or at least alleviate the disorders and then executes the treatment.

How the therapist evaluates speech and plans a treatment is not relevant to this thesis. Nor are the actual treatments used by the therapist to remediate or alleviate the effects of the disorder. We are interested in the feedback used by the therapist to indicate to the client that the client's speech is incorrect. People with speech disabilities often have poor auditory memories and find it hard to monitor their own speech. In order to correct their speech the clients are required to do various vocal exercises and repeat them over and over again. Initially they have to rely on their therapist to provide them with feedback about the correctness or incorrectness of their speech. Tape recorders are also used.

In the previous sections we have seen how various acoustic features of speech can be shown visually. In addition, from these visual displays many segmental and suprasegmental speech features can be identified (e.g. intonation patterns, stressed syllables, fricative sounds). The advent of the electronic age saw the beginning of the development of visual speech aids for use in clinics. The advent of microprocessor technology has increased the scope and potential of visual speech aids. The next chapter reviews the various visual speech aids that are available. After that the thesis focuses on our visual speech aid.

CHAPTER 2

VISUAL SPEECH AIDS

2.1 INTRODUCTION

The last chapter showed that various acoustic features of speech could be displayed visually (see sec 1.4). Pitch, loudness, phonation, resonance, voiced/unvoiced distinction, quality of sounds or words, timing, duration and stress are some of the many speech features that can be displayed. Any deviation from the accepted social norm in any of these features will result in unnatural sounding speech. Transforming speech from an acoustic signal to a visual signal can, therefore, provide feedback, additional to the aural feedback, about speech features of an utterance. This feedback can then be used to help the speech impaired improve their speech. This chapter is a review of visual speech aids and the different ways speech features have been visually presented. No attempt will be made to discuss how the different speech aids calculate the speech features.

In this thesis the aids that are discussed are called visual speech aids rather than visual speech and language aids. The aids are intended to be used for the remediation of speech and language at the phonological level. Spoken language has four levels; the phonological level (the sound system), the syntactic level (the grammatical arrangement of words and morphemes in a sentence), the semantic level (the meaning of words and utterances) and the pragmatic level (what does the speaker accomplish through making the utterance?) (Nation and Aram, 1977). Thus whilst the aids are used in language therapy, it is only at the first level. They provide no information at the syntactic, semantic and pragmatic levels of a language. Since the aids are to be predominantly used in speech therapy we have called them speech aids rather than speech and language aids.

Visual speech aids have been developed over the last 120 years. Up to thirty years ago the information provided about speech features was in some form of a "current value" response. A "current value" response is when the display results from the calculated or measured value of a speech feature obtained from one segment of speech. There is no record of the past values of the speech feature on the display. A voltmeter is an example of a "current value" display. The 1960's brought the advent of oscilloscopes with long persistence screens. The developers of the visual speech aids utilised this technology and developed aids which had displays which showed a record of temporal variations of a speech feature (for example loudness contours). Prior to the existence of the oscilloscope it would have been possible to obtain a record of the temporal variation of a speech feature using a chart recorder. However I have found no record of a visual speech aid which used one.

The coming of computer and microprocessor technology in the 1970's and 1980's

had a big impact on visual speech aids. Over the last 20 years many multi-featured speech aids have been developed.

In this review non-computer based visual speech aids will be briefly discussed in sec 2.2. In the section following (sec 2.3), 11 computer-based visual speech aids will be discussed. Particular emphasis will be placed on the speech features which are displayed and how the information is presented.

In this chapter and throughout this entire thesis displays which show variations in sound intensity will be called loudness displays. Those which show variations in fundamental frequency will be called pitch displays. It is slightly misleading to name the visual display of an acoustic characteristic after its perceptual correlate. However they are the names typically used in the literature and this is the nomenclature we will adopt.

2.2 NON-COMPUTER-BASED VISUAL SPEECH AIDS

Visual representations of the speech signal have existed for over 130 years. In 1859 Scott and Koenig developed the phonautograph, a device which traced sound waves on to smoked paper (Miller, 1934, p71-72). This device, however, was not used as a visual speech aid. Later, in 1862, Koenig went on to develop the Manometric flame. In this device "the flame of a burning gas jet vibrates in response to the variations in pressure in a sound wave" (Miller, 1934, p72). This research was leading up to the development of the telephone by A. G. Bell in 1876. Bell himself also saw the potential of applying some of the principles of telephony to create visual speech, for the teaching of the deaf to speak. In 1874 he used a "current value" visual display to provide feedback on the speech wave of a deaf pupil (Pickett, 1972).

By the 1920's, devices were used in speech training which used meters and lights to indicate pitch and inter-oral pressure (Boothroyd, 1992). Hudgins in 1935 wrote a paper entitled "Visual Aids in the Correction of Speech" for the Volta Review. In it he laid down three guidelines for the development of visual speech aids. He stated that the visual pattern must be simple so a child could understand it, the apparatus must be easy to operate even in the class room, and the aid must present the visual pattern while the child is speaking (Pronovost, 1967). All these points are as valid now, in the development of visual speech aids, as they were when they were made nearly 60 years ago. In 1944 the Bell Telephone laboratory combined a filter bank sound analyser with instantaneous phosphorescent screens to display a spectrogram. This device was called the visible speech translator (Pickett, 1972). It was tested extensively with deaf speakers.

By the 1960's many visual speech aids were available. The advent of the long persistence oscilloscope screen meant that aids were developed in which visual patterns could be made from the speech signals and retained on the screens for a period of time. These aids displayed such things as pitch contours (Stark *et al.*, 1968; Pickett and Constam, 1968; Risberg, 1968), loudness contours (Stark *et al.*, 1968; Pickett and Constam, 1968), Lissajous figures (Pronovost, 1963; Pronovost, 1967), sounds plotted on the first formant (F_1)- second formant (F_2) plane (Cohen, 1968; Pickett and Constam, 1968) and spectrograms (Risberg, 1968; Stark *et al.*, 1968).

Meters were still being used as visual indicators of speech features in aids in the 1960's, for example the s-indicator and nasalization meter (Risberg, 1968). In these

aids a needle would deflect if an [s] sound or a nasal sound was detected. Borrild (1968) described a speech training apparatus which had pitch and loudness meters. Several aids had specially built visual displays, usually a grid of lights. The Kamplex Visible Speech Apparatus, for example, had 10 neon tubes. It gave a spectral plot and was used as a vowel and consonant indicator (Searson, 1965). Thomas' F_1 - F_2 display is another example of a display based on a system of lights (Thomas, 1968; Thomas and Snell, 1970). It used a 12×12 matrix of neon lights to indicate the F_1 - F_2 formant plane. Different sounds lit up lights in different regions of the display.

All of the above aids were purpose-built, the hardware being such that the aid could perform only one function. Some of these aids could plot different aspects of speech (for example pitch contours, loudness contours and spectral content) on the same display medium, usually an oscilloscope screen. However each speech feature was calculated using specialized hardware circuitry.

This section only briefly outlined the non-computer-based visual speech aids. For a more detailed discussion the reader is referred to the references given. In addition, Pronovost (1967), Pickett (1972) and Lippmann (1982) present good reviews on the subject.

2.3 COMPUTER-BASED VISUAL SPEECH AIDS

The advent of digital technology and the computer vastly increased the potential and capability of visual speech aids. The computer screen is a much more versatile display medium than those used in visual speech aids of the past. In addition it does not require the constant and careful adjustments that were required for the oscilloscope screens (Nickerson and Stevens, 1973). The high resolution screen and the large colour palettes have allowed the speech features to be displayed in many novel ways. Loudness can be indicated by the size of a balloon, for example, instead of by the position of a deflected needle. Speech features can now be calculated via software algorithms rather than via hardware circuitry. Instead of having a purpose-built circuit for each speech feature, all the features can be extracted using digital signal processing hardware. This means that one computer-based aid could display many different features of speech. Each feature is calculated by a different software algorithm.

The first time a computer-based speech therapy aid was documented was in 1973 when Nickerson and Stevens asked the question: "Teaching Speech to the Deaf: Can a Computer Help?" Since then a large number of computer-based speech aids have been developed. In this section we will review eleven aids. The aids are: the **BBN aid** (Kallikow and Swets, 1972; Nickerson and Stevens, 1973; Stevens *et al.*, 1975; Boothroyd *et al.*, 1975; Nickerson *et al.*, 1976); the **Cambridge aids** (Crichton and Fallside, 1974; Fallside and Brooks, 1976; Brooks and Fallside, 1976; Fallside and Brooks, 1977; Bristow and Fallside, 1978; Bristow *et al.*, 1981; Brooks *et al.*, 1981; Bate *et al.*, 1982; Gulian *et al.*, 1984); the **VSA aid** (Povel, 1974a; Povel, 1974b; Povel and Wansink, 1986; Povel and Arends, 1991; Arends *et al.*, 1991); the **SSD aid** (Stewart *et al.*, 1976; Maki *et al.*, 1981); the **ISTRA aid** (Osberger *et al.*, 1978; Osberger *et al.*, 1981; Osberger *et al.*, 1982; Kewley Port *et al.*, 1987a; Kewley Port *et al.*, 1987b; Watson *et al.*, 1989; Kewley Port *et al.*, 1991); the **Yamanshi University aid** (Shigenaga and Kubo, 1986); the **ETSI Telecomunicacion aid** (Pardo, 1982; Aguilera *et al.*, 1986); the **IBM Speech Viewer** (Denoix, 1984; El Beze, 1986; Adams *et al.*, 1989; IBM, 1988); the

John Hopkins aid (Mashie *et al.*, 1985; Bernstein *et al.*, 1986; Bernstein *et al.*, 1988; Ferguson *et al.*, 1988; Mashie *et al.*, 1988) and the **Visible Speech aid** (Software Research,).

Many of these aids are in modular form. The therapist or client selects which module they want, depending on which speech feature requires remediation. Some of the aids have several modules for remediation of one speech feature, while others have one module per feature. The 11 aids will not be reviewed separately. The review will be structured according to the speech features the aids are intended to remediate. Particular attention will be paid to the manner in which the speech features are displayed. Each aid in the review will be referred to by its name as given in the preceding paragraph. No references will be given in the review. The information presented was all obtained from the series of references in the preceding paragraph.

There are two types of speech display in computer-based speech aids, as there were with non-computer based ones. The first type of display shows a record of how the speech feature varies with time. It will henceforth be called a time-plot. An example of a time-plot display is the spectrogram. Most of the modules in speech aids which display speech features on time-plots have two on the screen. One time-plot is for the target and the other is for the attempt. Unless specifically mentioned in this review, all the modules which display speech features on time-plots have two time-plots on the screen.

The second type of display provides no reference to time. The displays are obtained from the current value of a speech feature calculated or measured from a segment of speech. These displays will henceforth be called current-value-plots. An example of a current-value-plot is an s-indicator.

Visual displays for the remediation of suprasegmental features are discussed first followed by the displays for the remediation of segmental features.

2.3.1 Pitch Correctors

The different ways the eleven speech aids indicate pitch features of speech fall into four main groups. Two of the methods involve current-value-plots, the other two time-plots. In the first method, pitch is presented on a one dimensional display. The height of a marker is proportional to the pitch. The marker can move in the vertical but not horizontal direction. Three visual speech aids have a pitch module of this type. Pitch is represented by the height of the adam's apple in a cartoon character in the BBN aid module, the height of a character's nose in the VSA module and the height of the "mercury" in a thermometer in the IBM Speech-Viewer module. All these pitch displays are current-value-plots.

The second method of displaying pitch variations is similar to the first. Once again the height of a marker is proportional to the pitch, but this time the marker moves across the screen at a fixed rate. Four visual speech aids have modules which use this method. In the Ball Game module in the BBN aid, the client must try and get a ball through a hole in the wall at the left hand side of the screen. The client uses their pitch to control the height of the moving ball. In the VSA module the client uses their pitch to control the height of a moving head to get the head "to eat" as many fish as possible. The fish are at different heights on the screen. The tunnel game in the SIRENE aid involves the client using their pitch to move a marker through a tunnel of

varying height. The vertical position of the marker is controlled by the pitch. Finally, in the pitch skill building module of the IBM Speech-Viewer, variations in pitch guide a mobile marker through a series of obstacles (ghosts and treasure chests). Whilst the markers in the above four modules move horizontally across the screen, no trace is made of the path traversed. Thus all four modules are current-value-plots.

The most common way to display pitch variations is as a contour on a fundamental frequency vs time-plot. Five of the visual speech aids have modules that display pitch in this manner: the BBN aid, the SIRENE aid, the ETSI Telecomunicacion aid, the IBM Speech-Viewer and the Visual Speech Aid. All of these pitch modules have two time-plots on the screen, one for the target contour and one for the attempt.

The fourth method of displaying pitch is used in the John Hopkins University aid. It is a game using pitch contours, the F_0 Puzzle game. In the game the therapist requires the client to maintain their pitch within a certain frequency range. The therapist draws the frequency limits on the time-plot on the screen. If the client's trace remains within the limits then they are rewarded with a piece of a puzzle. Once the client has collected all the pieces of the puzzle they then get to unscramble the puzzle.

The pitch correctors with current-value-plots are clearly aimed at children. All are game like, for example, using pitch to change the height of a character's nose or to guide a marker through a tunnel. The modules would be good for learning pitch control. Other than that, however, they would have very little use, current-value-plots showing no record of the temporal variations in pitch. This means it would be difficult to indicate the pitch range, average pitch, pitch variation and pitch transition to the client using current-value-plots. All of these can be shown on the time-plots. The pitch time-plots could also be used to indicate rhythm and duration. These features could not be shown on the current-value-plots. The time-plot of pitch contours is a very informative display. To make the time-plots appealing to children a pictorial reward could be built into the module, similar to the F_0 puzzle game in the John Hopkins aid for example.

2.3.2 Loudness Correctors

Loudness variations are displayed either on current-value or time-plots. The loudness variations on the current-value-plots are shown in a variety of ways. Four of the modules use size to indicate loudness. Two of the speech aids, the BBN aid and the IBM Speech-Viewer, have modules in which size of a cartoon character mouth is proportional to the loudness of an utterance. As the vocal intensity increases the mouth gets bigger. The IBM Speech-Viewer has a second module in which the size of a balloon is proportional to the intensity. Based on a similar principle there is a module in the VSA aid in which the distance between a cartoon character's hands is proportional to the vocal intensity. As the intensity increases the hands get further apart.

The John Hopkins University aid has three modules which display vocal intensity on current-value-plots. The first module, the intensity game with immediate feedback, involves using vocal intensity to control the height of a balloon. Loudness is divided into three levels, soft, conversational and loud. These levels are indicated on a pole divided into three equal sections. The top section is red, the middle section is green and the bottom is blue. If the balloon is level with the red section the client is speaking loudly. If it is level with the green level the client is speaking at a conversational

level. Finally if the balloon is level with the bottom blue section, the client is speaking softly. The second module in the John Hopkins aid is the intensity game with delayed feedback. This is very similar to the first module; once again there is the balloon and the tri-coloured pole. The height of the balloon is controlled by vocal intensity. The client is required to articulate at a selected vocal intensity in this game. The intensity is indicated by a hand pointing to the target level on the pole. If the client articulates with the correct vocal intensity, the balloon will rise to the correct level. As an additional reward an animated cartoon sequence appears on the screen, for example a dancing rabbit.

The third vocal intensity module in the John Hopkins aid is the intensity game with limited feedback. In this game the client is required to articulate at a specified vocal intensity. This required intensity is indicated by the colour of a balloon. The colours of the balloon are red, green and blue. As with the previous two modules the colours correspond to loud, conversational and soft levels of intensity respectively. If the client articulates at the correct intensity they are rewarded with an animated cartoon series. The SIRENE aid has a module which is similar to the above module but is simpler. In the SIRENE module a drawing appears on the screen if the loudness of an utterance exceeds a pre-set level.

There were five aids which had modules which displayed loudness contours on an intensity vs. time-plot. These aids were the BBN aid, the ETSI Telecomunicacion aid, the ISTRa aid, the IBM Speech Viewer and the Visible Speech aid. In the modules of the IBM Speech Viewer and Visible Speech aids, the region between the loudness contour and the base line in the plots is shaded in.

The loudness correctors with current-value-plots are game-like, similar to the pitch correctors. Similarly, the same criticism of the current-value-plots for the pitch correctors also applies to those for the loudness correctors. Whilst the current-value-plots would be good for loudness control they have very little other use. It would be difficult to show loudness variations, loudness range, loudness transitions and average loudness on the current-value-plots. On time-plots of loudness contours, however, these trends would be easily seen. In addition rhythm and duration of speech could be seen on the time-plots but not on the current-value-plots. To make the time-plots more appealing to young children the module could have a game component included. Something similar to the F_0 Puzzle game in the John Hopkins aid, for example, or a visual reward like an animated cartoon sequence actuated at the discretion of the therapist.

2.3.3 Sustained Phonation and Voicing Correctors

Four of the speech aids have modules which are used to promote phonation, in the form of either sustained or short repeated vocalizations. The VSA aid has a module in which a man will walk along a road provided the client is vocalizing. A similar module is found in the Visible Speech aid. In this module the client gets a duck to walk down a pre-set path by producing a sustained phonation. In the Sustained Phonation Game of the John Hopkins aid sustained phonation shades in a rectangle of variable length. The length of the rectangle is set by the therapist. Interruption in vocalization will result in sections of the rectangle not being shaded in. If the client maintains phonation to shade in the entire rectangle, they are rewarded with an animated cartoon sequence.

The John Hopkins Aid has a second module which promotes repeated short vocal-

izations. In this module the client uses repeated short vocalizations to move a bird across the screen. The bird moves a step with each vocalization. A worm also moves across the screen at a fixed rate, independent of the client's ability to phonate. The aim is to get the bird to hop to the other end of the screen before the worm gets there. If the bird beats the worm then it picks up the worm in its beak and flies off. The speed of the worm is set by the therapist. The IBM speech viewer has a module based on a similar principle. To move a train down the track the client must produce a series of short vocalizations.

The IBM Speech Viewer has two other modules concerned with encouraging vocalization. The most elementary module is designed purely for sound awareness. When a sound is detected there is a continually changing multi-coloured pattern on the screen. If no sound is detected the screen is blank. The third module indicates voicing by the bow tie on a clown. The tie becomes covered in polka dots if voicing is detected.

Six of the aids have modules which are used to promote awareness between the difference in producing voiced and unvoiced sounds. The Voicing Skill module in the IBM Speech-Viewer involves the client using voiced and unvoiced sounds to guide a balloon over mountainous terrain and make it land. When a sound is voiced the balloon flies high. When a sound is unvoiced the balloon flies low. Four of the aids have a module in which voicing is indicated on a time-plot. These aids are: the BBN aid, the ETSI Telecomunicacion's ISOTON aid and the Visible Speech aid. If voicing is detected a trace appears on the time-plots, otherwise the screen is blank. The SIRENE aid also has a module which plots voicing on a time-plot. However in this module it is the contour of the energy in a given frequency range which is plotted. This frequency range is selected by the therapist. Thus the module detects voicing or unvoicing depending on the selected energy range.

The John Hopkins Aid has a module which gives a time-plot of the air flow through the glottis. This was measured by a pneumotachograph. This module can show the initiation and termination of phonation. It can also be used for the voiced /unvoiced distinction. There is a module in the VSA which measures voice quality by analysing glottal characteristics. Voice Quality is indicated by the position of a marker relative to one of two heads, one of the heads having a happy face and the other a sad face. The closer the marker is to the happy head the better the voice quality is. There are two voice quality plots on the screen, one for the therapist and the other for the client.

Similar to the Pitch and Loudness correctors, there are some Sustained Phonation and Voicing Correctors which have current-value-plots and some have time-plots. Once again the current-value-plots are more game like than the time-plots. However unlike the Pitch and Loudness correctors the current-value-plots for the Sustained Phonation and Voicing Correctors are very useful. Having a simple response on the screen as a result of phonation enables the client to be aware of when they phonate. The ability to phonate on command can be a difficult task for the speech-impaired child. The time-plots of voicing can also be used to show rhythm and duration in speech.

2.3.4 Articulation Correctors

Some of the articulation correctors are designed only to help with the articulation of phones (isolated sounds). Other correctors are designed to help with the correct pronunciation of any sound, word or phrase.

Seven of the speech aids had modules which were specifically to be used for the production of isolated vowels. The vowel corrector in the VSA aid display provides clients with instant feedback about which particular vowel they are articulating. The display of the vowel corrector is a two-dimensional space in which the coordinates are related to the first and second formants. Each vowel occupies its own region in the two-dimensional space. These regions are outlined on the display. The position of a marker in the display indicates which vowel is being articulated. The SIRENE aid has a similar module.

Speech recognition techniques were used in three vowel corrector modules of the aids reviewed. In the vowel corrector module in the Yamanshi University aid the client is requested to pronounce a specific vowel. A smiley face appears if they pronounce the vowel correctly. If the vowel was pronounced incorrectly a sad face appears and the client is told which vowel the computer deduced they were saying. The IBM Speech viewer has two modules for help with vowel pronunciation, both using speech recognition techniques. In the first module the client must guide a marker through a maze. The marker can move in four directions: up, down, left and right. The movement of the marker is controlled by correctly pronouncing one of four vowels, one vowel for each direction. In the second module the correct pronunciation of a vowel by the client enables a monkey to climb a tree.

The BBN aid has two modules designed for the monitoring of vowel production. The first module contains a time-plot of the first formant frequency (F_1) vs time. The height of the tongue is inversely related to F_1 (Ladefoged, 1975). Thus a plot of F_1 vs. time can be used for distinguishing between high and low vowels (see table 1.3). The second module contains a time-plot of the second formant frequency (F_2) vs. time. This module is used for distinguishing between front and back vowels. There is a relationship between the frontness of the tongue and the second formant. This is not as strong, however, as that between tongue height and F_1 (Ladefoged, 1975).

Three of the speech aids had modules in which reconstructions of the vocal tract shape were used to aid vowel production. All three modules reconstruct the vocal tract shape using linear prediction analysis. This method of reconstruction is discussed in detail in sec 8.3.2. In one of the Cambridge Aids, the Computer Vowel Trainer (CVT), the log area curves of the vocal tract vs distance were plotted on the screen. Each vowel has its own shaped curve. There were two traces on the screen, one being the target curve and the other the client's attempt. In the right hand corner of the screen there was a cartoon of a bear. This bear smiled if the client's attempt was a good match with the target.

The modules of ETSI Telecomunicacion's SAS aid and the Yamanshi University aid displayed the reconstructed vocal tract on the background of a mid-sagittal cross-section of the head. The movement of the jaw is shown as well as the vocal tract shape. In the SAS aid there are two vocal tract shape plots superimposed on the mid-sagittal cross-section of the head. One plot is for the target shape, set by the therapist, and the other is the client's attempt. If the client's attempt makes a good match with the target trace a cartoon character in the corner of the screen smiles (this is similar to the CVT aid). The module in the Yamanshi University aid also gives an indication of lip rounding.

Three aids have modules for practising the pronunciation of any isolated sound. The SIRENE and IBM Speech Viewer aids both have a module in which the short-time

spectrum of a sound is displayed on a frequency vs. energy plot. The BBN aid had a module called the vertical spectrum. The display in the module is obtained from the short-time spectrum, and is vase-like in shape. Each sound had its own shape. In the module there were two displays on the screen, one for the target shape and the other for the attempt by the client. This module also provided an indication of whether a sound was voiced or unvoiced. When voiced sounds are detected, horizontal lines appear across the shape. These lines disappear for unvoiced sounds.

The Fricative Training aid (the other Cambridge aid) displays the occurrence of frication, voicing and silence in an utterance against time. This aid has two plots, one for the target and the other for the attempt. The display is of a band evolving left to right across the screen. It is either black (indicating silence), grey (indicating voicing) or white and black chequered (indicating unvoiced frication).

There are two aids, the SIRENE and the SSD, which have modules in which the spectrogram of the utterances is plotted. These two modules can be used to remediate misarticulation when it occurs in a sound, word or phrase. The module in the SSD aid plots a grey scale 1.5 or 0.75 second spectrogram. The choice of the duration of the spectrogram is selected by the user.

The Articulation Correctors which were designed to be used for practising the correct articulation of a specific phone were all current-value-plots. Some of these were game-like (for example the Maze module in the IBM speech viewer) and most displays were of the actual calculated speech feature (for example the vowel corrector in the VSA aid). It is difficult to use Articulation Correctors based on current-value-plots to indicate the correct pronunciation of sounds in words and sentences. However they are very useful for practice of the phones in isolation and for learning to distinguish between different phones.

There is a need for both time-plots and current-value-plots in Articulation Correctors. Current-value plots enable specific sounds to be practised in isolation. Once these sounds are correctly pronounced the time-plots can then be used to investigate how effective the carry-over of the correct pronunciation is in words, sentences and spontaneous speech.

The response of the vowel corrector module in the Yamanshi University aid for incorrect pronunciations is questionable. The aid displays a smiley face if a vowel is pronounced correctly and a sad face if the vowel is pronounced incorrectly. Having an animated display regardless of the correctness of the pronunciation means that the client is being rewarded what-ever they do. It would be better if the screen was exceedingly dull if the pronunciation was incorrect. Thus the motivation would be to get an interesting screen, i.e. an animated display.

2.3.5 Other Speech Correctors

Several of the aids had modules to be used in speech therapy which did not fit into any of the categories discussed so far.

Six of the modules in the BBN aid represent several speech features simultaneously. In the cartoon face game the height of the adam's apple in the throat was related to the pitch, the size of the mouth was related to the vocal intensity and if any [s] or [z] sounds were correctly uttered then [s] or [z] respectively would appear in a cartoon conversation bubble on the screen. There were another five modules which showed

several speech features concurrently on a time-plot. The modules gave plots of: pitch and loudness; voicing and loudness; voicing and nasalization; and voicing, loudness and nasalization. The nasalization was measured by an accelerometer at the nose. All the voicing in the BBN aid was measured by an accelerometer at the throat.

The clown face module in the IBM speech viewer was a simpler version of the cartoon face game in the BBN aid, showing the two speech features of voicing and intensity simultaneously. If voicing was detected the bow tie on the clown became polka dotted and the size of the clown's mouth was related to vocal intensity, the louder the sound the bigger the mouth.

The Visible Speech aid had a module called the rhythm detector, which displayed voicing on a time-plot. The time-plots of the pitch and loudness contours, mentioned in sec 2.3.1 and 2.3.2, can also be used for displaying duration, stress and rhythm.

The final visible speech aid which will be reviewed is the ISTR A aid. This aid is used for helping with articulation, but it is quite different from the aids discussed in sec 2.3.4. Most of the articulation modules in the aids reviewed so far have displays which result from visually representing calculated speech features, e.g the vertical spectrum in the BBN aid. In some of the modules in the aids, speech recognition techniques were used. There was a visual response if a specific isolated sound was recognised by the computer, for example the Maze module in the IBM speech viewer. However the modules were only used to practise the articulation of isolated sounds. The ISTR A aid compares the client's attempts at pronouncing words with their best attempts at the words thus far. The aid utilizes speech recognition technology in five of its modules. The therapist must continually update the "best attempt" templates.

The bar graph and target modules give a response on the screen for every utterance. In the bar graph module each utterance is evaluated as to whether it was a good, fair or bad production, this being indicated by the level of the bar. In the target module the participants try to produce an utterance which is as good as the stored "best attempt". A marker on a target indicates how "good" the client's production is. The closer the attempt is to the template, the closer the marker is to the bull's eye in the target. In the Baseball and Race modules, points are scored if the client's attempt is within a certain preset threshold of the stored "best attempt" template.

The fifth module is call Moonride. In this module a rocket will move a bit closer to the moon each time a selected word or utterance is pronounced within a pre-set threshold of the stored template a specific number of times. This number is set by the therapist. The higher the number the more elaborate the graphic reward is when the rocket lands on the moon. There is a final module of the ISTR A aid which plots the loudness contours on a time-plot; this has already been mentioned in sec 2.3.2

The ISTR A aid is used for remediation of all vowel errors (in words or in isolation), the consonant errors of omission, insertion, and, in most cases, substitution and finally, in the remediation of suprasegmental features except those involving pitch.

The literature on the ISTR A claims that it is an effective visual speech aid. However it has a serious drawback. Once a client has improved their speech to the level of the best response a new "best attempt" must be recorded. The decision as to which utterance of the client is the current "best attempt" is made by the therapist. The ISTR A would not enable the client to learn to assess their speech from interpreting the visual displays. When practising alone the client could only improve their speech to the level of the pre-stored "best-attempt". If their speech improved beyond the pre-stored

“best-attempt” the client would have no way of knowing.

2.3.6 The Hardware Components and User-Interface Software

Up till now no mention has been made of the hardware of the eleven aids. There is little point in going into great detail on this subject as the aids in this review extend over a 19 year period. There have been many advances in computer technology in this time. Comparing the hardware capabilities of the aids would be like comparing different machines. The BBN aid (developed around the early to mid 1970's), for example, was built on a DEC PDP 8/E. In contrast the modified VSA (developed in the early 1990's) was built on a Commodore Amiga 2000. The VSA itself was built on various computers, its development beginning in the 1970's. Each computer reflected the technology of the time.

Regardless of the advances in technology the basic hardware components of the speech aids were similar. All the aids in the review processed speech in real-time. Each had a microphone as the primary acoustic sensor. Three aids had additional sensors. The BBN aid, the VSA aid the John Hopkins aid had a sensor which measured the glottal movements. These were a throat accelerometer for the BBN aid and an electroglottograph for the VSA and John Hopkins aids. The BBN aid and the John Hopkins aid each had a third sensor, a nose accelerometer and a pneumotachograph (a device for measuring air flow) respectively.

All the aids had a control unit, usually a key board. Some had a control panel of dials and switches, for example the BBN aid and the SSD. Each aid had a processing unit in which the speech was analysed. Finally each aid had a display medium, a video monitor. The early aids, such as the BBN aid and the Cambridge aids had a monochromatic video monitor with low resolution. As time went by the screen resolution increased as did the size of the colour palette. This meant the most recent aids have detailed graphics. The displays of the modified VSA aid are good examples of this.

The modules within the BBN aid, the Cambridge aid and the ETSI Telecommunication aids were all run by independent programs. The modules in the VSA, ISTR A, SIRENE, IBM Speech Viewer and the John Hopkins University aids were menu driven. This made it very easy for the user to select which module they wanted to use. The SSD, whilst computer-based, was essentially purpose-built to display spectrograms. No information was given as to the user interface of the Visible Speech aid or the Yamanshi University aid.

The VSA and the John Hopkins aids both had two versions. One version of the aid was for use of the speech therapist and client in the speech clinic, the other was for the client at home. The aids for use in the clinic contained administrative data about the clients, as well as the speech modules. The John Hopkins aid for the clinic had three acoustic sensors, the microphone, the electroglottograph and the pneumotachograph. The home unit only contained the microphone. The therapist was able to load pre-selected exercises into the home unit of the VSA. The ISTR A aid also had a module in which administrative data about the clients could be placed, but there was only one version of the aid.

Using acoustic sensors, in addition to the microphone, in a visual speech aid has the advantage that different speech features can be displayed. The nose accelerometer, for instance, means nasalization can be detected. However there are draw-backs in

using too many acoustic sensors. Most people initially feel self-conscious speaking into a microphone. If they have to use other acoustic sensors as well, the client may feel very uncomfortable; this would hinder their efforts to improve their speech.

2.4 CONCLUSION

Throughout this review we have seen how aspects of speech such as pitch, loudness, phonation, resonance, the voiced/unvoiced distinction, the “goodness” of the production of sounds or words, timing, duration and stress can be displayed visually. In addition there are many different ways of displaying the information. Some of the aids displayed the speech features on current-value-plots, others on time-plots and others utilized both methods of display. The current-value-plots were usually in some game format in which variations in the speech feature corresponded to the changing of the height or width of some object, for example the height of a balloon corresponding to vocal intensity. Some of the current values were of a visual representation of an actual feature of speech, for example the vocal tract reconstruction plots. The time-plots were all used to show how a particular speech feature varied over time. No games were made out of the time-plot displays. Some modules had pictorial rewards which appeared on the screen, for example the F_0 puzzle game in the John Hopkins aid.

Most of the displays of visual speech aids required a certain amount of interpretation by the users to assess the “goodness” of production. This might be relating balloon size to vocal intensity or learning to “read” a spectrogram. The displays from modules which employed speech recognition techniques did not require interpretation of the quality of the production, as it was done by the computer. However the client still had to relate a movement to a sound, for example an [a] sound moving the marker to the left in the Maze module of the IBM speech viewer.

Two speech therapists with whom I have worked, Evelyn Terrice and Kate McNabb, have said that one of the most liberating aspects of the visual speech aids for the speech-impaired is that they can “see” their mistakes. The speech-impaired do not usually notice that they are speaking incorrectly. It is usually some-one else who informs them of the errors they produce. Since it is more empowering to identify one’s own mistakes, the speech impaired were more motivated to improve their own speech.

It is questionable whether an aid which assesses the correctness of an utterance by matching it with a pre-stored template (for example the ISTR A aid) would provide the clients with a sense of control. It is the therapist who decides what the ‘best-attempt’ template will be. When practising alone the client can only improve their speech to the level of the “best-attempt”. If they actually make a better attempt at the utterance (i.e a more correct pronunciation) they will have no way of knowing from the display and, worse, the attempt will be indicated as being poorer than the reference. However if the displays resulted only from a calculable speech feature the client would be able to see for themselves if their speech was improving. It is true that these sorts of displays do require a certain amount of interpretation by the user (client or therapist), but this can be taught.

This chapter has outlined several different visual speech aids, it has discussed the types of speech features the aids displayed and it has described how the features are presented.

In the following chapter, chapter 3, the Computer Aided Speech Therapy Tool

(CASTT), a computer-based aid developed in the Department of Electrical and Electronic Engineering at the University of Canterbury will be compared to the speech aids reviewed in this chapter.

CHAPTER 3

THE COMPUTER AIDED SPEECH THERAPY TOOL (CASTT)

3.1 INTRODUCTION

In the previous chapter we reviewed many visual speech aids, and saw the different ways speech features could be displayed visually. In this chapter a visual speech aid, developed by the Speech Group in the Department of Electrical and Electronic Engineering at the University of Canterbury, will be presented. The aid is called the Computer-Aided Speech Therapy Tool (henceforth called the CASTT). The details about the hardware and software used in the CASTT will be described in chapter 4, as will the theory on how the speech features are calculated.

The CASTT project began in 1982 when Professor R. H. T. Bates was given a research grant by the Telethon Trust (International Year of the Disabled) to provide technical advice and assistance to the New Zealand speech therapy community. A real-time signal processing system, incorporating the digital signal processing chip the TMS32010 within an IBM-PC XT, was developed by Steven Turner for his Master's project (Turner, 1986). In the years 1986 to 1987 three postgraduate students of the speech group (Andrew Elder, Tracy Clark and Turner) worked on a real-time computer-based speech therapy aid project. David Moon, an undergraduate student, also worked on the project for his final year project in 1986. Five potential speech analysis modules were developed (Elder *et al.*, 1987), (Bates *et al.*, 1987). At the end of 1987 I took over the real-time computer-based speech therapy aid project and became the sole person on the project.

Since 1988 the CASTT has been extensively evaluated by 15 speech therapists in the Canterbury region. These evaluations were an important component of the CASTT development. They are outlined and discussed in chapter 5. Through the evaluations many changes were made to the modules in the CASTT. In addition two more modules have been added since I took over the CASTT project. The changes made to the modules will be briefly discussed in chapter 5. The purpose of this chapter is to present the CASTT in its current form.

3.2 THE CASTT

The CASTT is a visual speech aid designed to remediate errors in suprasegmental aspects of speech, in articulation and in phonation. It has seven modules. Three of these are for the remediation of suprasegmental aspects of speech. These modules are the Voice Pitch Tracker (a pitch corrector), the Loudness Monitor (a loudness corrector) and the Concurrent Loudness and Pitch module (a pitch and loudness corrector).

The CASTT has three modules which are articulation correctors, the Spectrogram

module, the Fricative Monitor and the Vocal Tract Shape module. The Spectrogram module can be used for practising the articulation of sounds in isolation, in words and in sentences. The Fricative Monitor is used for practising the articulation of fricatives. It is intended for sounds uttered in isolation. The Vocal Tract Shape module is used for practising the articulation of isolated vowels. It can be used for both monophthongs and diphthongs. The final module in the CASTT is used for practising phonation. It is called the Sustained Phonation Monitor.

Each of the displays in the CASTT is essentially a plot of the calculated values resulting from an analysis of speech. There are no game-like plots, such as representing loudness variations by the size of a balloon (see sec 2.3.2) or using pitch variations to guide a marker through a tunnel of varying height (see sec 2.3.1). The lack of games in the CASTT was partly because the methods of displaying the speech features in all the modules were inherited from the initial people involved in the project. However it was also due to the author's dislike of relating variations in speech features to an unrelated action (for example what has loudness got to do with the size of a balloon?). The Sustained Phonation module in the CASTT is the only module which has a game component. If the client's phonation attempt was to the satisfaction of the therapist, the therapist could actuate an animated cartoon sequence. There were five different sequences. The computer selected the sequences randomly. This meant the client had no idea which sequence would be the reward.

The displays of the Voice Pitch Tracker, the Loudness Monitor and the Concurrent Loudness and Pitch module were all time-plots. This, as discussed in sec 2.3.1 and 2.3.2, is a very powerful display medium for pitch and loudness characteristics. From the plots the range, variation, average values of the pitch and loudness can be seen. In addition the displays can also be used for errors in rhythm or duration.

The articulation correctors in the CASTT have displays which are either time-plots or current-value-plots. In the review of visual speech aids (see sec 2.3.4) both the current-value-plot and time-plot articulation correctors were recommended. The correctors with current-value-plots can be used for practising isolated sounds. Once these sounds are mastered the correctors with time-plots can be used to practise the carry-over of the sound into words, sentences and spontaneous speech. The Spectrogram module has a time-plot display. The Fricative Monitor and the Vocal Tract Shape module have current value displays. The final module of the CASTT, the Sustained Phonation module has a time-plot display.

Five of the modules in the CASTT are similar to at least one of the modules mentioned in the review in sec 2.3. However none of the aids in the review have the same combination of modules as are in the CASTT. In addition most of the modules in the CASTT contain features which none of the modules in the aids reviewed possess.

Having introduced the CASTT, each of the seven modules will now be presented.

3.2.1 The Voice Pitch Tracker Module

The Voice Pitch tracker displays a record of the variations of the pitch vs. time. Figure 3.1 gives the screen display of the Voice Pitch Tracker. It can be seen that the module has two display graphs. The top display graph is for the target contour, the bottom for the attempt. Having two display graphs means that comparisons can be made between the target and attempt contours. For more accurate comparison it is

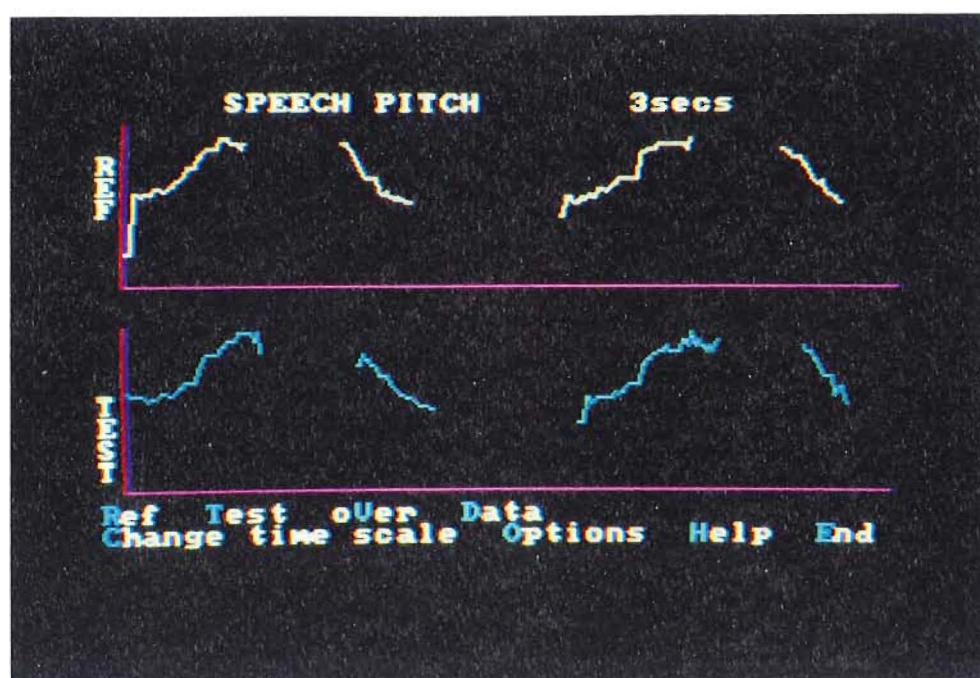


Figure 3.1. The Screen display of the Voice Pitch Tracker for the utterance "now? now! now? now! ".

possible to overlay the contour of the bottom display graph on the contour of the top graph. It is also possible to directly superimpose the contour which would appear in the bottom display graph onto the top display graph as the person is talking.

The superimpose option means the Voice Pitch tracker can be used for pitch control. The therapist can set the target contour and then the client must attempt to vary their pitch so their contour varies in the same manner. Many of the aids reviewed in sec 2.3.1 had time-plots which displayed pitch contours. However none of these aids had the capacity to overlay or superimpose the attempt contour onto the target contour. The Voice Pitch Tracker can also be used to show pitch variation, pitch range, pitch transitions, average pitch, rhythm and duration.

The pitch contour in figure 3.1 is of the utterance " now? now! now? now! ". The variations in pitch can be seen clearly in the display. In the module it is possible to call up a listing which gives the average fundamental frequency, the highest frequency, the lowest frequency and the frequency range in the contour of either display graph. The pitch contours can be stored on disk. At a later date the contours can be recalled and restored on the display graphs. This function has many uses. For example a client can see how their speech has improved or a therapist could store several target contours for the client to practise imitating. Both the John Hopkins aid and the VSA aid, discussed in sec 2.3, have the facility to store the displays of the target or the attempt.

The Voice Pitch Tracker also has an audio feedback option. The option enables the last three seconds of speech input to be replayed through an audio speaker. By relating the pitch contour to the replayed speech the client is provided with additional feedback about their utterance. None of the aids in the review (see sec 2.3) mention any modules which have an audio feedback option.

Finally an on-line help is available to the users of this module.

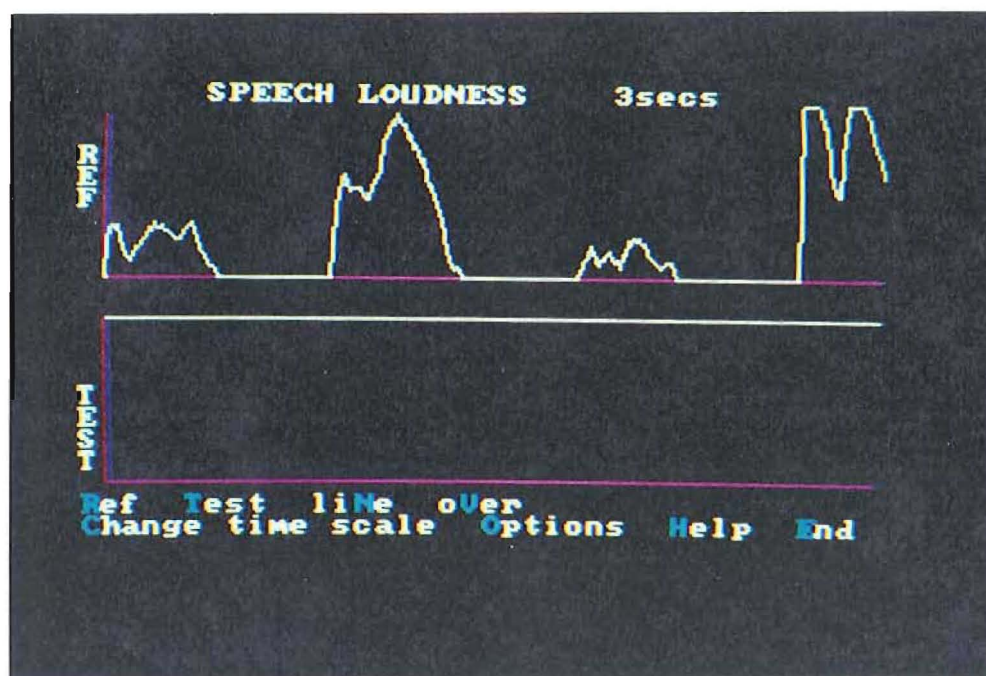


Figure 3.2. The Screen display of the Loudness Monitor for the utterance “hullo hullo hullo hullo”, spoken softly and loudly alternatively.

3.2.2 Loudness Monitor

The Loudness Monitor displays a record of the variations of the loudness (vocal intensity) vs. time. Similar to the Voice Pitch tracker, the Loudness Monitor has two display graphs. The top display graph (see fig 3.2) is for the target loudness contour, typically set by the therapist. The bottom display graph is for the attempt. It is possible in the Loudness Monitor, as in the Voice Pitch Tracker, to overlay the contour in the bottom display graph onto the top display graph. The superimpose option is also available. None of the aids reviewed in sec 2.3.2 had the overlay or superimpose options. The Loudness Monitor, like the Voice Pitch Tracker module, had the audio feedback option, on-line help and the ability to store contours on disk.

The loudness contours in fig 3.2 are of the word “hullo” spoken softly and loudly alternatively. The variation in loudness can be seen very clearly. There is a level indicator on the bottom display graph, a moveable horizontal line. The height of the line provides a loudness target for the client. If the client speaks too softly, for example, they would be required to articulate again so that the height of the loudness contours could exceed that of the line. This feature and the superimpose option mean that the module can be used for loudness control. The Loudness monitor can also show loudness variations, loudness range, loudness transition, average loudness, rhythm and duration.

3.2.3 The Concurrent Loudness And Pitch Module

The Concurrent Loudness and Pitch module displays a record of both the variations of pitch and loudness vs. time. The contours are displayed on two display graphs. The pitch contours are displayed on the top graph (see fig 3.3). The loudness contours are

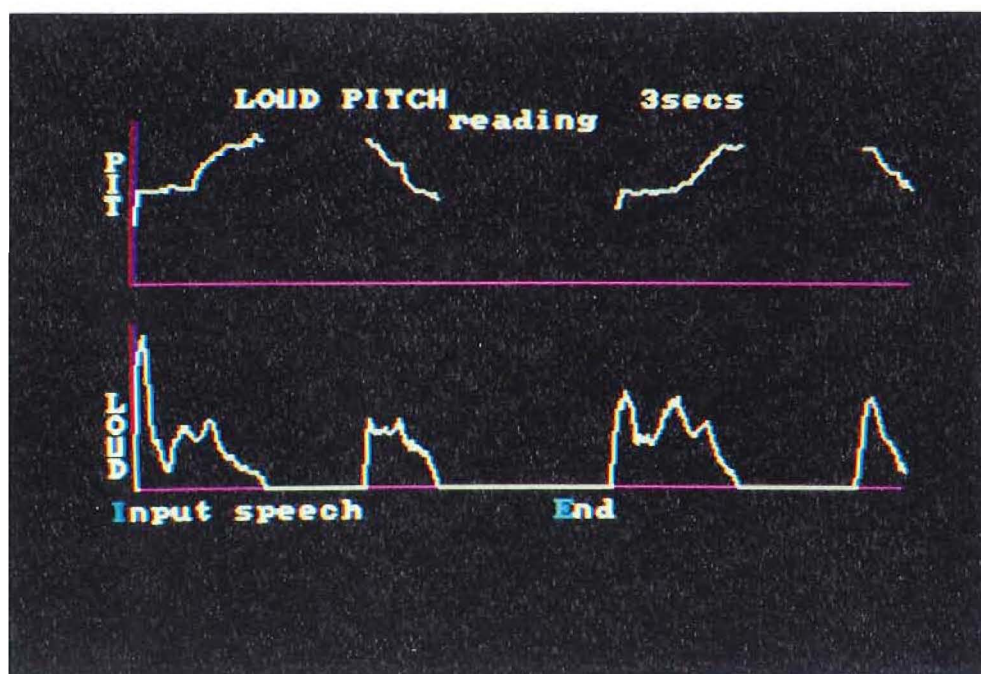


Figure 3.3. The Screen display of the Concurrent Pitch and Loudness module for the utterance "hullo? hullo! hullo? hullo!".

displayed on the bottom one (see fig 3.3).

Figure 3.3. gives the display for the utterance "hullo?, hullo!, hullo?, hullo!". The variations in pitch and loudness can be seen clearly. There was a module similar to this CASTT module in the BBN aid in the review (see sec 2.3.5). The BBN aid was the only aid in the review which displayed different speech features concurrently on a time-plot. In the review most of the modules which displayed concurrent speech features had current-value-plots.

The Concurrent Loudness and Pitch module was designed specifically for showing intonation to the hearing-impaired. It is difficult for the hearing-impaired to differentiate between variation in loudness and variation in pitch. The module can be used to show which speech feature is being varied.

3.2.4 The Spectrogram Module

The Spectrogram Module plots real-time spectrograms from speech input via a microphone. The spectrogram is obtained by calculating a 256 point FFT (see sec 4.5.3). The spectrogram is a pattern of four different colours, each of which represents an energy band. Figure 3.4 gives the screen display for the utterance "me bee". The differences and similarities between the two words can be clearly seen. The spectrographic pattern for the [i] portion in the two words is quite similar. However the pattern corresponding to the [m] and [b] portions of the words, is quite different.

The Spectrogram module, like the spectrogram module in the SIRENE aid and the SSD, has two display graphs. This means comparisons between the target and the attempt can be made. The Spectrogram module can be used for practising the articulation of sounds in isolation, in words, sentences and in spontaneous speech.

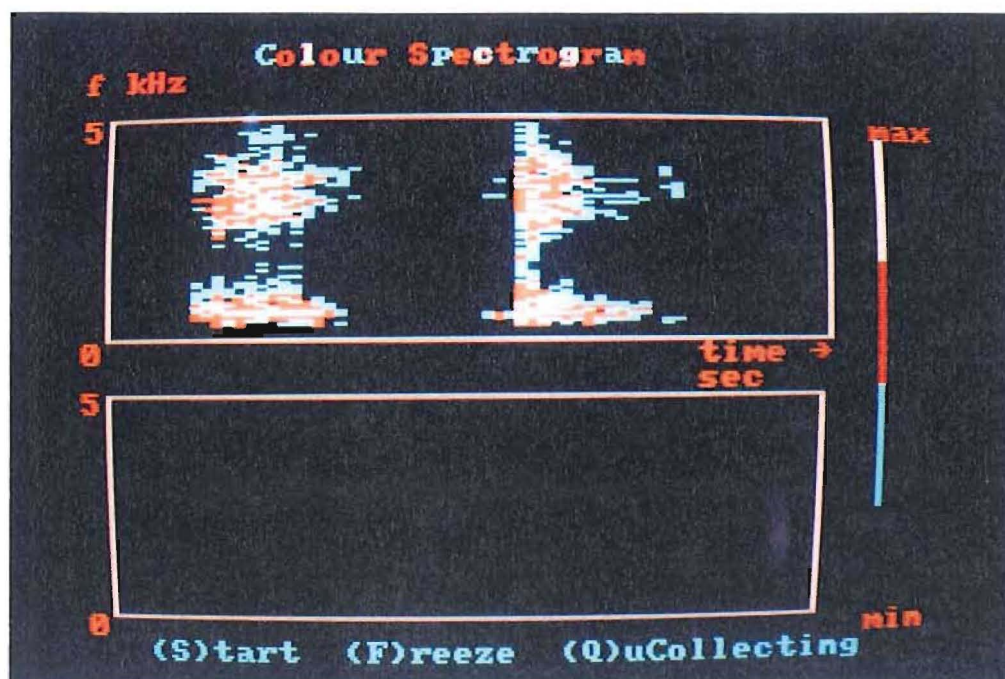


Figure 3.4. The screen display of the Spectrogram module for the utterance "me bee".

3.2.5 The Fricative Monitor

The Fricative Monitor is designed to indicate when it detects frication. If frication is detected a longitudinal bar appears on the screen. If no frication is detected there is no response on the screen. It is possible to adjust the sensitivity of the monitor to frication. Figure 3.5 is an example of the screen display for the utterance [ʃ]. The longitudinal bar can be seen, indicating that [ʃ] is a fricative (as expected).

In the modules of the aids reviewed in sec 2.3 there is no module similar to the Fricative Monitor. The Fricative Monitor display was designed to be similar to the display of an S-indicator. The displays of both are very simple. There is only a response if an [s] or a fricative is detected.

3.2.6 The Vocal Tract Shape Module

The Vocal Tract Shape Module is used for practising the articulation of vowels. The reconstruction of the vocal tract shape, calculated from speech in real-time, is displayed on a mid-sagittal cross-section of the head. The jaw movements are also shown on the display.

The module is similar to the ETSI Telecommunication's SAS aid and a module in the Yamanshi University aid (see sec 2.3.4). Unlike those two modules, however, there are two mid-sagittal cross-sections of the head on the screen (see figure 3.6). This means comparisons between a target vocal tract shape and an attempt can be made. For closer comparison it is possible to superimpose the target shape on the other head. This enables the client to attempt to match the target shape.

The SAS aid had two vocal tract shapes in its display. However both were permanently overlayed on the same mid-sagittal cross-section of a head. The vocal tract

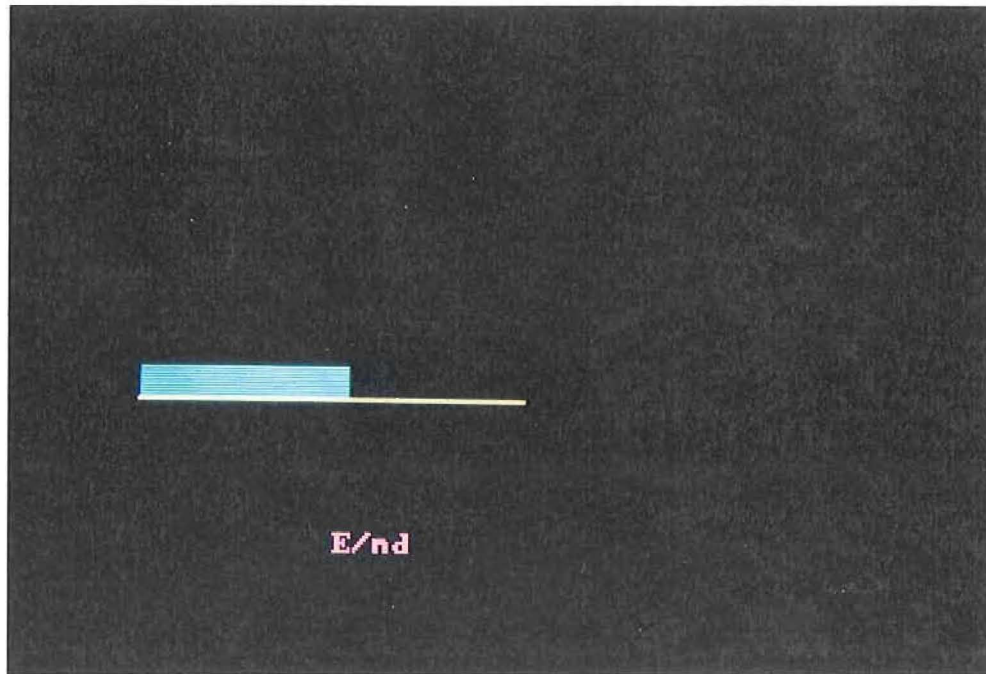


Figure 3.5. The screen display for the Fricative Monitor module for the utterance [ʃ].

shape module in the CASTT has the advantage that the client or the therapist can decide whether they want the target shape to be overlayed or not. The overlayed vocal tract shape can be erased when desired.

The vocal tract shape in the CASTT module is updated every 40.4 ms. The vocal tract shape assumes a default neutral shape if no sound is detected by the microphone for over 0.404 seconds. This means the display is not distracting when the module is not being used.

3.2.7 The Sustained Phonation Module

The Sustained Phonation module is used for practising phonation. Whenever the client phonates, a colourful display appears on the screen. The module can also be used for practising sustained phonation. The pattern on the screen remains fairly constant for a sustained isolated sound. The Sustained Phonation Monitor module has a game component. The clients are required to sustain phonation of a specified sound for a certain length of time. After the child client has finished phonating, the following question appears on the screen “has the pattern reached the end without changing ? Y/N”. If the answer is “yes”, the client is then rewarded with 1 of 5 pictures, chosen randomly by the IBM-PC. If the answer is “no”, the message “Have another go” appears on the screen. It is the speech therapist who makes the decision as to whether the client has been successful or not.

It was a deliberate decision to have a dull screen (the message “Have another go”) if the client was not successful in sustaining their phonation. If the screen had been vaguely interesting, for example an unhappy face, then the client would still have been rewarded, even though they had not been successful. The intention of this module was

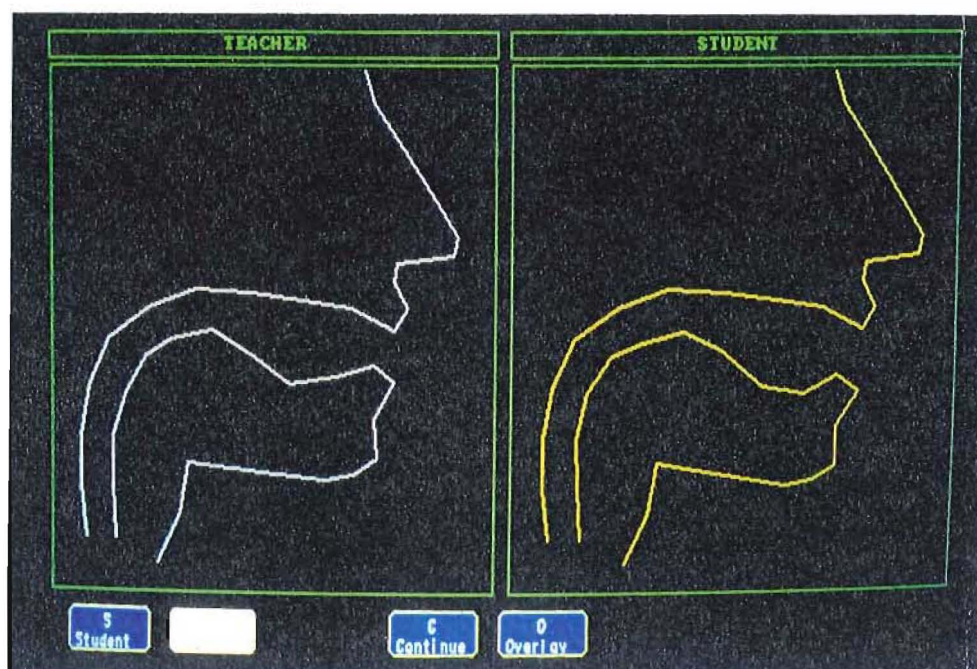


Figure 3.6. The screen display for the Vocal Tract Shape module showing the estimated vocal tract shapes of the sound “ahh” in both the TEACHER and STUDENT windows.

to provide positive reinforcement only.

An example of the screen display for the Sustained Phonation Module and one of the pictorial rewards is given in figure 3.7 and figure 3.8 respectively.

None of the aids in the review (see sec 2.3) had a module similar to the Sustained Phonation module in the CASTT.

3.2.8 The CASTT Shell Software Package

The CASTT aid was menu driven. This is similar to the VSA, ISTR, SIRENE, IBM Speech Viewer and John Hopkins aids (see sec 2.3.6). The CASTT software package was designed so people with very little familiarity with the IBM-PC operating system could use the aid. The package calls up a supervisory menu from which each of the separately executable modules can be selected. In addition, the package provides users with instruction on how to run the CASTT software packages. For example if floppy disks are required, a message indicates which floppy disk to use and how to load it.

The modules are selected by pushing an appropriate key from the keyboard. All the modules are listed on the screen menu. The highlighted letter indicates the key to be pressed to select the desired module (see figure 3.9). Throughout all the CASTT modules the same method of executing instructions or commands was used. All the commands available in the modules were given in menus on the screen.

3.2.9 The Hardware Of The CASTT

The CASTT consisted of an acoustic sensor (a microphone), a control unit (a keyboard), a processing unit (a purpose-built speech-processing board and an IBM-PC XT), audio

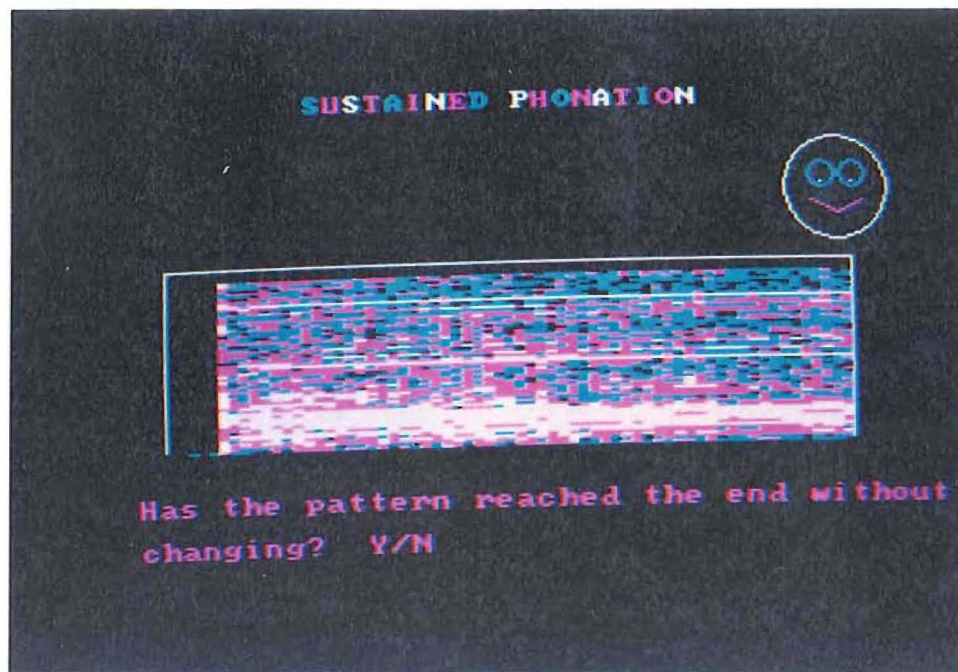


Figure 3.7. The screen display for the Sustained Phonation module for a sustained sound.

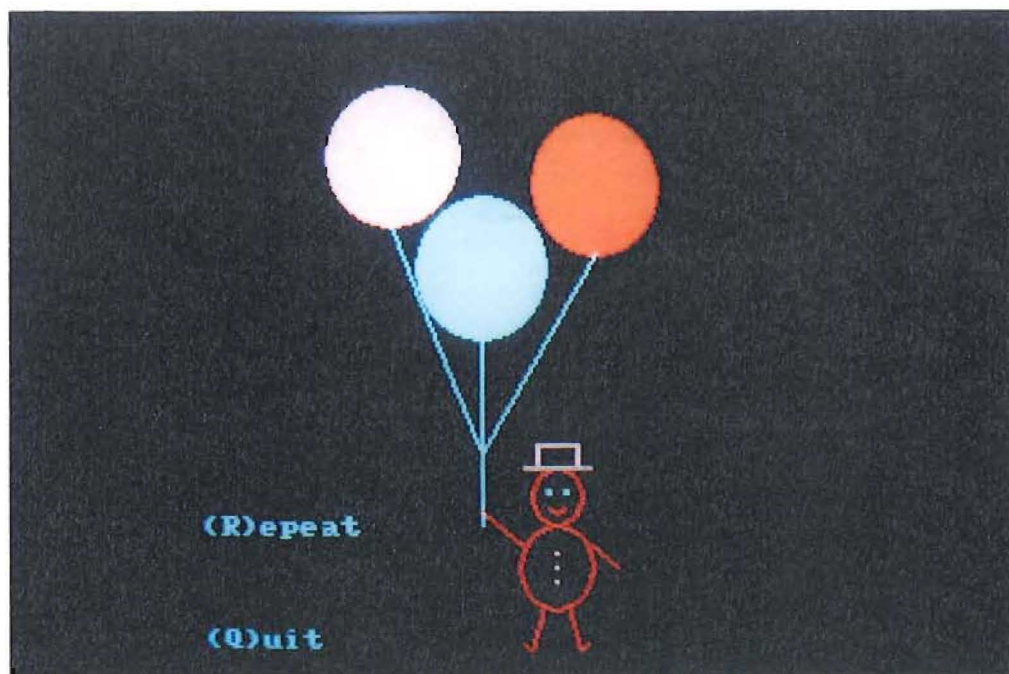


Figure 3.8. An example of one of the five pictorial rewards of the Sustain Phonation module.

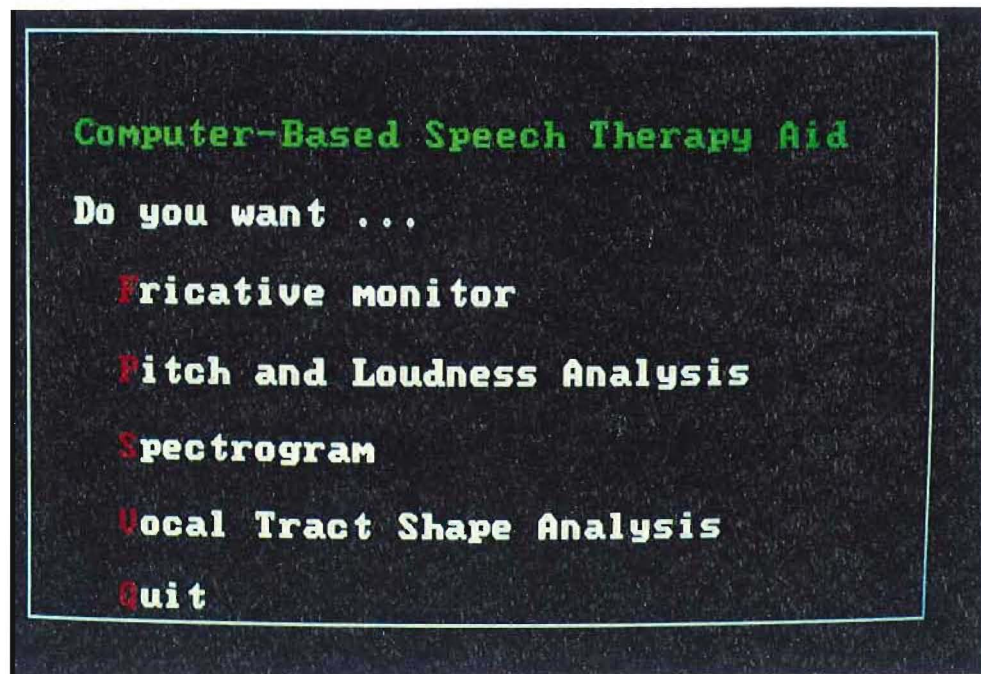


Figure 3.9. The speech analysis module menu.

feedback (a loudspeaker) and a display medium (a video monitor). These are the same basic hardware components of all the aids discussed in sec 2.3.6. The hardware of the CASTT will be discussed in greater detail in sec 4.2.

Having described the seven modules of the CASTT, the signal processing aspects utilized to obtain the displays in the CASTT will be discussed in the following chapter.

CHAPTER 4

SIGNAL PROCESSING FOR THE CASTT

In the last chapter the seven speech analysis modules of the CASTT were described. For each module a description of the displays was given. The displays are essentially plots of calculated values resulting from analyses of speech. Speech analysis in the CASTT is performed by digital signal processing, which requires the signals to be represented as a sequence of numbers, i.e. digitized.

Though there are seven speech modules in the CASTT, only five different methods of speech analysis are used. This chapter will outline all five methods. It will also outline the analysis method utilised by a possible new CASTT module. However, before that, the digitizing of speech and some of the associated problems will be presented. In addition, the hardware and software of the CASTT which enables the digitizing and processing of speech will be discussed.

4.1 DIGITIZED SPEECH

All the signal processing for the CASTT is performed on digitized signals, but speech is a continuous signal. To obtain a digitized signal from a continuous signal, the continuous signal is sampled and then quantised. The resulting sequence of numbers is the digitized signal. In this section we will first discuss sampling and then quantisation.

The relationship between sampled speech, $s(n)$, and continuous speech, $s(t)$, can be represented mathematically as:

$$s(n) = s(t)\delta(t - nT) \quad n = 0, \pm 1, \dots, \pm \infty \quad (4.1)$$

where T is the sampling interval ($\frac{1}{T}$ is the sampling frequency), $s(n)$ is the n th sample of the continuous signal $s(t)$, and $\delta(t)$ is the Dirac impulse function. It is possible to completely reconstruct the original signal, $s(t)$, from its sampled version, $s(n)$, provided $s(t)$ is bandlimited and was sampled at least at the Nyquist frequency. The Nyquist frequency is twice the frequency of the maximum frequency component of $s(t)$. Since $s(t)$ is bandlimited there will be a maximum frequency component. If $s(t)$ is sampled at less than the Nyquist frequency then aliasing occurs. All the frequencies of $s(t)$ greater than half the sampling rate “seemingly take on the identity of a lower frequency” in its sampled version, $s(n)$ (Rabiner and Schafer, 1978, p26). Therefore it is not possible to reconstruct the original signal $s(t)$ from $s(n)$, and the reconstructed $s(t)$ is distorted.

The highest possible frequency component in speech is not known exactly. Most of the energy of speech sounds lies below 5 kHz, although for some sounds, e.g. fricatives, there is significant energy up to 10 kHz. Harrington (1988) has suggested there is significant energy up to 16 kHz for the sounds /f/, /v/, /θ/, /ð/, /h/. For this reason it is necessary to bandlimit speech to a known frequency range before the speech is digitized

(Owens, 1993, p20). This is done by passing the speech through a low-pass filter. The sampling rate of speech is normally somewhere between 6 and 20 kHz, depending on the application (Owens, 1993, p21).

Passing speech through a low-pass filter does not completely reject frequencies in the stop band of the filter. These frequencies are typically attenuated by 60 to 70 dB (Owens, 1993, p21), rendering them insignificant. To ensure that frequencies greater than half the sampling frequency are attenuated by at least 60 dB, the cutoff frequency of the low-pass filter is chosen to be somewhat less than half the sampling frequency. For example in digital telephone systems the cut-off frequency for the low-pass filter may be 3.4 kHz with a sampling rate of 8 kHz (Witten, 1982, p52).

Digital sampling of a waveform is discrete not only in time but also in amplitude. The digital encoding of the amplitude of a sampled signal $s(n)$ is known as quantisation (Clark and Yallop, 1990, p231). The amplitude is represented as a binary number. In this thesis the digitized signal is identified by square brackets, thus $s[n]$ is the n th digitized sample of the continuous signal $s(t)$. The number of bits representing the amplitude determines the accuracy of the digitized signal.

Fry (1979, p94) states that the average intensity of speech at the conversational level is about 60 dB; this increases to about 75 dB for shouting and decreases to about 35 to 40 dB for very quiet (though unwhispered) speech; all these measurements were measured at a distance of 3 feet and 0 dB was the threshold of audibility. Quantizing the amplitude to a 10 bit binary number is sufficient to represent an intensity range from 0 - 60 dB.

This section has discussed what is necessary to digitize a continuous signal, such as speech. The following section will discuss how the speech is digitized in the CASTT. In order for this to be done it is first necessary to present the hardware of the CASTT. The software of the CASTT will be discussed in the section after.

4.2 THE HARDWARE AND SOFTWARE OF THE CASTT

The hardware of the CASTT comprises a microphone, a keyboard, an IBM-PC XT and a purpose-built speech-processing board. All the digitizing of the speech in the CASTT was done on the purpose-built speech-processing board. This board, henceforth called the TMSboard, contains a TMS32010 digital signal processing chip, dual access RAM, an Analogue to Digital (A/D) converter, a Digital to Analogue (D/A) converter and supporting hardware. It was designed to achieve real-time speech processing (Turner, 1986). The board was developed by Steven Turner (1986) as part of his Master's project.

To digitize speech the continuous speech signals are input into the TMSboard from either an external amplifier or a standard microphone. The speech signal is bandlimited to a known frequency range, before it is digitized. To do this it is passed through a 6 pole low-pass filter which has a cutoff of 3.0 kHz. The roll-off of the filter response is such that by 5kHz the frequencies components are being attenuated by 27 dB, and by 10 kHz they are being attenuated by 60 dB. The speech is then digitized by an A/D converter at a 10 kHz sampling rate. The sampled amplitudes are 12 bit numbers. Quantizing the amplitude to be a 12 bit number means the intensity range for signals in the CASTT is 70 dB.

The energy in most speech sounds lies below 5kHz (Ainsworth, 1988, p54). At 3.0

kHz, the cutoff point of the anti-aliasing filter is going to have a noticeable effect on the accuracy of how some speech sounds will be represented in the modules of the CASTT. Recall that the displays of the CASTT are essentially plots of calculated values resulting from analyses of speech. Sounds which have high energy in the spectrum above 3.0 kHz, such as fricatives and stops, will have a significant amount of information removed from the signal by the anti-aliasing filter. This means the displays of the Spectrogram module cannot accurately represent these sounds. However provided all the sounds can be discriminated from the displays it does not matter that the representation is inaccurate. It is the premise of this thesis that providing speech errors can be distinguished by means of the displays of the speech aid, the aid will have remedial potential (this will be discussed at length in Chapter 6).

The 3.0 kHz cut-off of the anti-aliasing filter will not affect the accuracy of the pitch and loudness contours in the Loudness Monitor, Voice Pitch Tracker and Concurrent Loudness and Pitch modules. Nor will it detract from the accuracy of the Vocal Tract Shape module display. This module is only intended for the articulation correction of vowels. All the first three formants of vowels lie below 3.5 kHz (Rabiner and Schafer, 1978, p174). It is only the first three formants that are of primary importance in the representation of vowel sounds (Markel and Gray, Jr., 1976, p164). The 3.0 kHz cut-off of the anti-aliasing filter will affect the output of the Fricative Monitor module, as fricative sounds have significant amounts of energy in their spectrum above 3 kHz. These effects will be discussed later in sec 5.2.1.

Once the speech has been digitized it can be processed in the TMS32010 digital signal processor, which resides in the TMSboard. All the plots displayed by the CASTT modules result from speech analysis performed in the TMS32010 digital signal processor. It was necessary to use a separate digital signal processing system (i.e. not the processor in the IBM-PC XT or compatible) for real-time speech analysis because the general purpose microprocessors available in the XT's were not fast enough. The TMS32010 is able to perform a 16 bit multiplication operation in 200ns, compared with 28 μ s for the Intel 8088 (the microprocessor in the IBM-PC XT's)(Elder *et al.*, 1987). It is essential that the CASTT is able to display speech features in real-time. One of Hudgins' (1935) guidelines for an effective visual speech aid was that the aid must present the visual pattern as the client speaks (Pronovost, 1967).

The CASTT was built on technology that evolved from the mid-1980's. Nowadays there are much more powerful IBM-PCs available than the IBM-PC XT. For example, the Intel 486 DX microprocessor can perform a 16 bit multiplication in 520-1040 ns (INTEL, 1993, p2-356 to 2-359). Whilst this is considerable faster than the Intel 8088 (the microprocessor on the IBM-XT) it is still at least 2.5 times slower than the TMS32010 digital signal processor so it would still be necessary to have a separate microprocessor to achieve real-time signal processing. It should be added that digital signal processing chips have also become much more powerful since the mid-80's. The TMS320C30, a top of the line digital signal processor, produced by Texas Instruments, can do a 32 bit multiplication in 60 ns.

4.2.1 Software

As a consequence of real-time analysis, speech is processed in sequential segments. The duration of the segments depends on the application of the speech processing. Speech

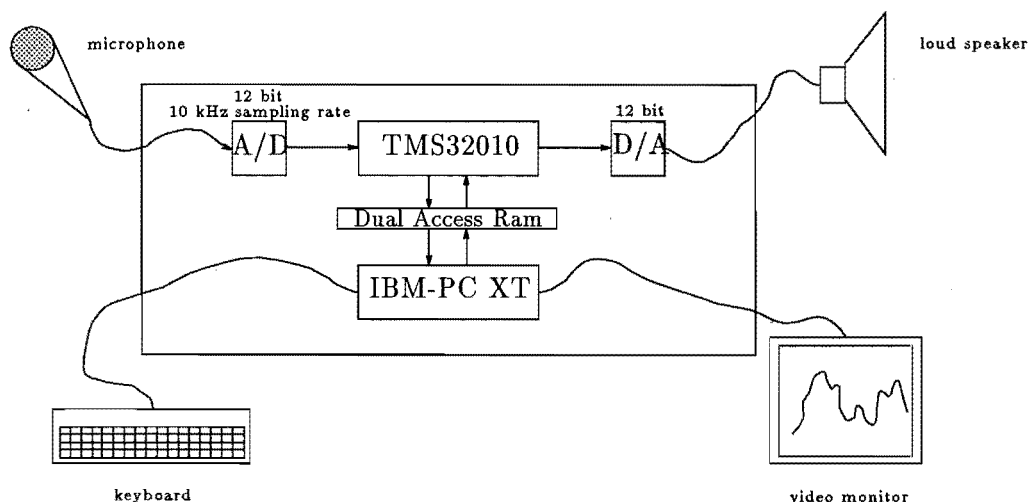


Figure 4.1. The real-time computer based visual speech therapy aid, the Computer Aided Speech Therapy Tool (CASTT).

analysis is essentially the process of estimating the time-varying features of speech. An updated value for a speech feature is calculated for each segment. Whilst the features of speech are time-varying, over a short period of time they can be considered to be non-varying. Due to physiological constraints, such as the rate of movement of the articulators, speech is considered to be a stationary signal over a 10-30ms interval in time. Hence for all the speech processing in the CASTT modules, the speech signal is processed in sequential segments of 10 to 30 ms. For example, in the Voice Pitch Tracker module in the CASTT a new pitch value is calculated for every 101 speech samples. This means that the pitch value is updated every 10.1 ms, since the sampling frequency is 10 kHz. Each of these pitch values is then plotted on the display plots. By drawing lines between consecutive pitch values, the pitch contours are obtained (see for example those in fig 3.1).

Once a segment of speech has been analyzed in the TMS32010, the results are passed to the dual-access RAM (see fig 4.1). This is memory that can also be accessed by the IBM-PC XT or compatible (henceforth the IBM-PC XT or compatible will be referred to as an IBM-PC). The IBM-PC collects the results from the shared memory and plots them on the video monitor. The graphics in the CASTT are controlled by a Colour Graphics Adapter (CGA) card. The CGA graphics card limits the resolution of the video monitor to 320×200 pixels. It also means that only four different colours can be used for displays. The design of the TMSboard enables the TMS32010 to perform uninterrupted speech analysis whilst the IBM-PC updates the displays (Turner, 1986).

The TMSboard can also convert digitized speech back into an analogue signal. The digitized speech, stored in the IBM-PC, is passed through the D/A converter on the TMSboard. The resulting analogue signal is then passed through a 2 pole 4 kHz low-pass filter to remove high frequency components caused by the sampling process (Turner, 1986). The signals can then be output to an external amplifier or a loud speaker. The CASTT has its own loud speaker. Figure 4.1 illustrates the basic hardware components of the CASTT.

4.2.2 Programming

All the instructions for the speech analysis on the TMS32010 are written in TMS32010 code. There is limited memory space in the TMS32010. Due to this, all the TMS32010 programs are stored in the IBM-PC and the applicable program is down-loaded into the TMS32010 when it is required. Each of the seven speech analysis modules in the CASTT consists of two executable programs, an IBM-PC executable and a TMS32010 executable. The two executable programs run in parallel. The IBM-PC executable down-loads the applicable TMS32010 code, collects the speech features which have been calculated by the TMS32010 and plots the information on the screen. Some additional processing is performed in the IBM-PC for the Vocal Tract Shape module.

The language used for the IBM-PC programs was not the same for each module. The IBM-PC software for the Loudness Monitor, the Voice Pitch Tracker and the Concurrent Loudness and Pitch modules was written in Microsoft Pascal. For the Fricative Monitor, Spectrogram and Sustained Phonation Module the software was written in Turbo Pascal and for the Vocal Tract Shape module it was written in Modula-2.

The software of the CASTT was written by more than one person. In addition to myself, Tracy Clark, Andrew Elder, Bill Kennedy, David Moon, Steven Turner and Brenda Satherley have all written software for the CASTT. I joined the CASTT project after all the above, except Brenda. Hence the software written by myself in the CASTT's modules was mostly adapting the pre-existing software. I have made substantial contributions to the codes for the Voice Pitch Tracker and Loudness Monitor modules, written all the IBM-PC code for the Concurrent Loudness and Pitch module, completely revised the IBM-PC and TMS32010 codes for the Vocal Tract Shape module in addition to rewriting the IBM-PC code for that module from Turbo Pascal to Modula-2. The latter was done to capitalise on the greater graphic capabilities of Modula-2 over Turbo Pascal. Modula-2 supported Meta-Windows, a windows-based graphics package. The version of Turbo Pascal available at that time, could not support Meta-Windows.

Having outlined aspects of digitizing speech and the hardware and software of the CASTT, I can now discuss the speech analysis methods used in the CASTT's modules. Thus the rest of this chapter is devoted to speech processing techniques. At the beginning of this chapter it was mentioned that, whilst there were seven speech modules in the CASTT, only five different speech analyses are performed. These are, loudness analysis, pitch analysis, spectral analysis, zero crossing analysis and vocal tract area analysis. All these methods are quite distinct from each other, there being no overlap in the analysis techniques. For each of the five methods of analysis, the theory behind the method will be presented along with the algorithm used to perform the analysis on the CASTT. The methods of speech analysis used in each CASTT module will also be given. Some modules use the results from only one method of speech analysis, some use the results from a couple of methods. The final section of this chapter will discuss the analysis method behind a possible new CASTT module, which would display Lissajous figures of speech. In order to display Lissajous figures, the speech signal must be phase-shifted. The phase-shifting algorithm and how the Lissajous figures are obtained will be discussed.

4.3 LOUDNESS ANALYSIS

The amplitude of any sound is proportional to the magnitude of the air pressure variations in the sound vibration (Clark and Yallop, 1990, p197). For sounds produced by humans, the greater the subglottal pressure, the greater the amplitude and the louder the sound appears. Thus loudness is proportional to the subglottal air pressure.

However loudness is not an acoustic measure. Rather, it is the perceptual correlate of the sound intensity, which in turn is proportional to the square of the acoustic amplitude and to the frequency of the signal. If there are two sounds of the same amplitude but one sound is at a higher frequency than the other, the sound with the highest frequency will appear the loudest. To obtain an estimate of the sound intensity it is common practice to calculate it from the amplitude of the signal using the root mean square (RMS) value of the amplitude, the frequency component being ignored. These RMS values are proportional to the sound intensity (Clark and Yallop, 1990, p200). The loudness contours displayed by the Loudness Monitor, (see fig 3.2) are plots of the RMS values of the amplitudes of digitized speech in the sequential segments of speech. Recall that, for all the modules in the CASTT, speech is processed in sequential segments (see sec 4.1).

For a single segment of speech containing N samples, the RMS amplitude value is calculated by:

$$RMS = \frac{1}{N} \sqrt{\sum_{n=0}^{N-1} s^2[n]} \quad (4.2)$$

where $s[n]$ is the n th digitized sample, $n = 0, 1, \dots, N - 1$, of the continuous speech signal $s(t)$.

4.3.1 The TMS32010 Root Mean Square Algorithm

In the TMS32010 RMS algorithm a new RMS value is calculated for every 101 samples of speech. This means that the RMS value is being updated once every 10.1 ms. A flow chart of the RMS algorithm is given in figure 4.2. It can be seen from the flow chart that the RMS value passed to the IBM PC is in fact 101 times the actual RMS value. There is no need to perform the division of (4.2) in the algorithm of fig 4.2, since the number of speech samples, N , will always be 101. This saves computation time. The relative difference between each consecutive RMS value remains the same regardless of whether the division is performed or not, and the shape of the contours would also remain unchanged.

There is no square root function in the TMS32010 assembler language to calculate the RMS value. An algorithm which calculates the square root of any number, developed by Hwang (1979) was used to perform this function in the TMS32010 RMS algorithm.

4.3.2 The Modules Which Use The TMS32010 RMS algorithm

There are four modules in the CASTT which use the TMS32010 RMS algorithm. The contours in the Loudness Monitor are calculated from RMS values of speech, as mentioned in sec 4.3. The loudness contours in the Concurrent Loudness and Pitch module are also calculated using the RMS algorithm.

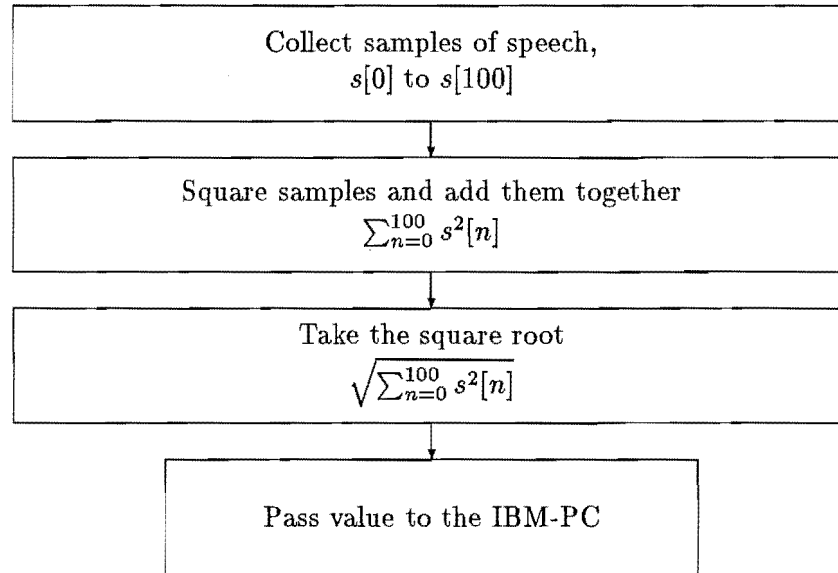


Figure 4.2. The TMS32010 RMS algorithm which calculates the RMS of a speech signal every 10.1ms and passes the value to the IBM-PC.

The plotting of contours in the Loudness Monitor, Voice Pitch Tracker and Concurrent Loudness and Pitch modules does not begin until the RMS value of the speech has exceeded a pre-set threshold. This means the plotting of the contours (pitch or loudness) only occurs once the user begins to speak. This gives users time to prepare themselves to speak. Hence the Voice Pitch Tracker module also utilises the RMS algorithm.

The final module that makes use of the TMS32010 RMS algorithm is the Vocal Tract Shape module. When the module is activated the vocal tract shape displays of both mid-sagittal cross-sections of the head are in neutral positions (see sec 3.2.6). The estimation of the vocal tract shape of the active head only begins when the RMS value exceeds a certain pre-set threshold.

The vocal tract shape estimation will stop if silence is detected. Silence is said to occur if the continually updated RMS values do not exceed the pre-set threshold for 0.404 s. If silence occurs then the vocal tract shape assumes the neutral position. The vocal tract shape remains in this position until the RMS value exceeds the pre-set threshold again.

4.4 PITCH ANALYSIS

It is possible to estimate the fundamental frequency of a person's voice from the time-domain signal. The frequency does vary, but a periodic component can be seen quite clearly in all the voiced portions of the speech signal. The calculation of the fundamental frequency of a speech signal is called pitch extraction. This title is not strictly correct since pitch is not an acoustic measure (as mentioned in sec 1.4). It is the perceptual

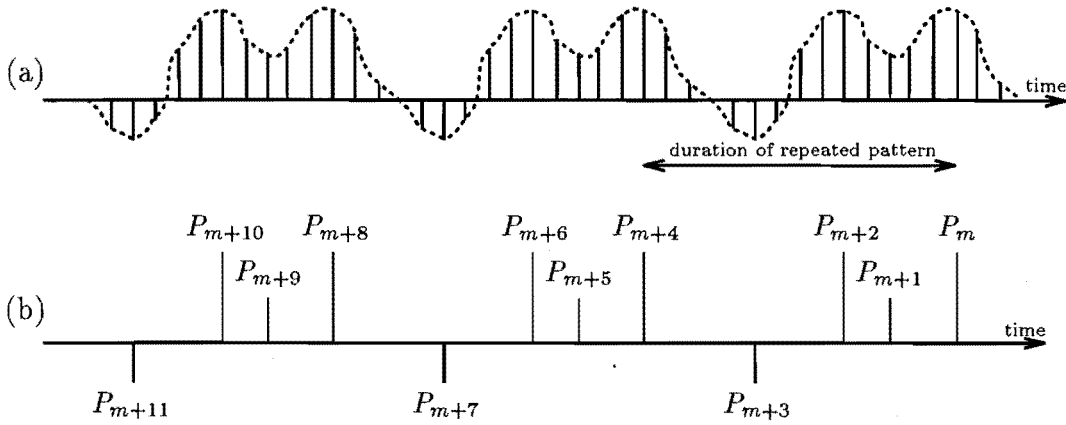


Figure 4.3. (a) A digitized periodic signal, where the dotted outline indicates the original analogue signal and the solid lines represent the amplitudes of the digitized values, and (b) the extrema of the digitized signal in (a).

correlate of the fundamental frequency. However pitch extraction is the title that has been adopted in common usage, so we will retain the nomenclature (as mentioned in sec 2.1).

There are many different algorithms for extracting pitch. Hess (1992) has discussed some of the many methods. This thesis will present only the method used in the Voice Pitch Tracker. This algorithm is a heuristic method developed by Brieseman (1984). The easiest way to explain Brieseman's algorithm is by way of an example. The repeated pattern in the periodic signal illustrated in fig 4.3(a) is easily recognisable by eye. Brieseman's algorithm is based on the notion of identifying the duration of the repeated pattern in a periodic signal. The duration of the pattern is the reciprocal of the pitch of the signal. The first thing Brieseman's algorithm does is locate all the extrema in the speech segment and calculate the duration between each pair of adjacent extrema. The extrema in the signal illustrated in fig 4.3(a) are plotted in fig 4.3(b). A repeatable pattern can also be seen in fig 4.3(b). It can be seen that the duration of this pattern is the same as the duration of the pattern illustrated in fig 4.3(a). Hence by identifying the duration of the repeated pattern in fig 4.3(b), the duration of the repeated pattern in fig 4.3(a) is also identified. Thus the pitch of the periodic signal can be calculated.

To get the first estimate of the pitch, the amplitude of each extremum in fig 4.3(b), starting with P_{m+1} , is sequentially compared with the amplitude of the most recent extremum P_m to find the first extremum which has an amplitude the same as or very close to P_m . The extrema are labelled backwards, with P_m being the most recent extremum and P_{m+11} being the first occurring extremum in the signal in fig 4.3(a). This is because Brieseman's algorithm has the pitch estimation beginning from the most recent extremum. It makes the mathematical notation simpler if the extrema are numbered back to front. The first extremum the algorithm will pick is P_{m+2} , since the amplitudes of P_m and P_{m+2} are very similar (see fig 4.3(b)). Thus the first pitch period estimate is the time lapse between P_{m+2} and P_m . Since we have supposedly identified the duration of the repeated pattern, it stands to reason that the amplitudes of P_{m+1} and P_{m+3} must be similar because they are the next extrema along from P_m and P_{m+2} respectively. However it can be seen quite clearly from fig 4.3(b) that the amplitudes of P_{m+1} and P_{m+3} are not similar, so the first estimation of the pitch period was wrong.

The next estimation of the pitch period would be the time lapse between P_m and P_{m+4} . The amplitudes of these two extrema are very similar. It can be seen in fig 4.3(b) that the amplitudes of P_{m+1} and P_{m+5} are also very similar (as are the amplitudes of P_{m+2} and P_{m+6} , P_{m+3} and P_{m+7} , etc). Thus the duration of the repeated pattern in fig 4.3(b), and hence fig 4.3(a) has been found. The pitch of the signal in fig 4.3(a) can be calculated from the reciprocal of the duration of the repeated pattern.

In more formal terms Brieseman's algorithm is as follows: For a segment of speech find all the extrema and the numbers of speech samples between extrema (remember we are dealing with digitized signals so the time between extrema can be calculated from the number of samples). The time, τ_a , between two consecutive extrema, say P_a and P_{a+1} , can be calculated from:

$$\tau_a = \frac{l_a}{F_s}, \quad (4.3)$$

where l_a is the number of speech samples between the extrema P_a and P_{a+1} and F_s is the sampling frequency. Suppose that in a segment of speech with q extrema, the amplitudes of the peaks are $P_m, P_{m+1}, \dots, P_{m+q-1}$, where P_m is the most recent extremum. Associated with those extrema are the inter-peak times $\tau_m, \tau_{m+1}, \dots, \tau_{m+q-1}$ (recall that τ_m is the time between the extrema P_m and P_{m+1} , etc). Next the smallest integer p is found such that

$$|P_m - P_{m+p}| < \varepsilon \quad (4.4)$$

where ε is some preset tolerance and p lies in the range $0 < p < INT(q/2)$, where $INT()$ is the integer function. If a value p satisfies (4.4) then the first estimation of the pitch period, T_0 , is

$$T_0 = \sum_{j=0}^{p-1} \tau_{m+j} \quad (4.5)$$

and the first estimation of the pitch is $\frac{1}{T_0}$ (Brieseman, 1984).

Once an estimate of the pitch has been made it is checked to see if it is a good estimation of the pitch, that is, has the duration of the repeated pattern been identified? Recall that, the extremum for which (4.4) holds is p extrema away from the most recent extremum P_m . Each pair of extrema in the speech segment which are p extrema apart are checked to see whether the difference in their amplitudes is less than the pre-set tolerance, ε . If all the differences are less than ε then the duration of the repeated pattern has been identified and a pitch estimate can be made. If any differences are greater than ε then the pitch estimate is considered to be wrong and a new one must be made. Starting again with the extremum P_m , the next value of p is found such that (4.4) is true. The pitch checking process, described above, repeats again until a new estimate is made. If no value of the pitch is found for a speech segment the pitch value calculated for the previous speech segment is carried forward as the pitch value (recall speech is processed in sequential segments).

There is a limit to the lowest pitch value Brieseman's algorithm can calculate. It is limited by the number of speech samples in the speech segment for which the pitch is calculated and by the sampling frequency. The algorithm requires that the speech segment must be large enough to accommodate the extrema of at least two pitch periods of some pre-determined maximum pitch period, so a period match can be found (Brieseman, 1984). In the Voice Pitch Tracker in the CASTT, the lowest pitch

value the module can estimate is 100 Hz. The value of the pre-set tolerance, ε , was suggested by Brieseman (1984) to be in the vicinity of 10% to 30% of the maximum extremum amplitude in the set of extrema from which the pitch period estimation is made. The value ε must be large enough to account for the envelope variations of the speech waveform and changing values of the extrema as the signal develops with time.

4.4.1 The TMS32010 Pitch Algorithm

The TMS32010 pitch algorithm broke up Brieseman's original algorithm into two parts. The first part, the peak detection, located the extrema of the signal. The second part, the pitch extraction, estimated the pitch period of the speech signal from the extrema. A new pitch period value was calculated for every 101 speech samples. The TMS32010 pitch algorithm calculates the pitch period of a speech segment and passes it to the IBM-PC. The pitch for the speech segment, which is merely the reciprocal of the pitch period, is calculated in the IBM-PC. It should be mentioned that the pitch can only be estimated for voiced speech, there being no periodic component in unvoiced speech from which the pitch can be estimated.

4.4.1.1 Peak Detection

The first part of Brieseman's algorithm detects the extrema in the speech segment. However, before any extremum detection occurs the speech samples are passed through a 17 tap finite impulse response (FIR) 700 Hz low-pass digital filter which utilises a Hamming Window (windowing signals is discussed in sec 4.5.2). The filter coefficients are generated by a program, written by R. Stephens (Turner, 1986), called FIR2. This program ran on the departmental VAX computer.

The filter serves two purposes. Firstly it smooths out high frequency noise on the waveform. Secondly filtering the speech reduced the amount of computation needed to calculate the pitch because minor peaks are removed. The effects of this are illustrated in fig 4.4. Figure 4.4 (a) illustrates the unfiltered version of a portion of speech and fig 4.4 (b) is the filtered version of the same speech portion. It can be seen that the pitch remains the same for both the unfiltered and filtered signal. However the filtered signal is a much smoother waveform than the unfiltered signal.

The peak detection method is based on the comparison of the orientation of slopes on the filtered speech waveform. Consider the consecutive filtered speech samples $s_f[n-1]$, $s_f[n]$ and $s_f[n+1]$, where $0 < n < N-1$ (it should be noted that though the extrema are labelled backwards from the most recent extremum, the digitized speech is labelled in the forward direction; thus for a speech segment comprising N samples the most recent speech sample is $s[N-1]$). Let $DIFF_NEW$ be defined as

$$DIFF_NEW = s_f[n+1] - s_f[n] \quad (4.6)$$

and let $DIFF_OLD$ be defined as

$$DIFF_OLD = s_f[n] - s_f[n-1] \quad (4.7)$$

If $DIFF_NEW$ and $DIFF_OLD$ have the same sign then there is no extremum at $s_f[n]$, as illustrated in fig 4.5(a). If $DIFF_NEW$ and $DIFF_OLD$ do differ in sign an extremum at $s[n]$ has been identified, as illustrated in fig 4.5(b).

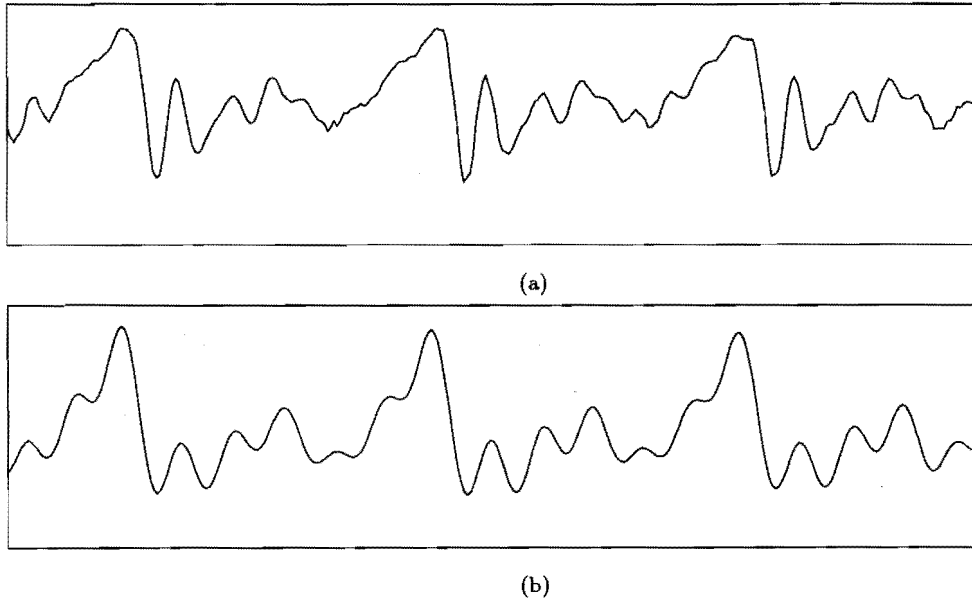


Figure 4.4. A segment of speech which is (a) unfiltered and (b) filtered by a 17 tap FIR 700 Hz low-pass digital filter.

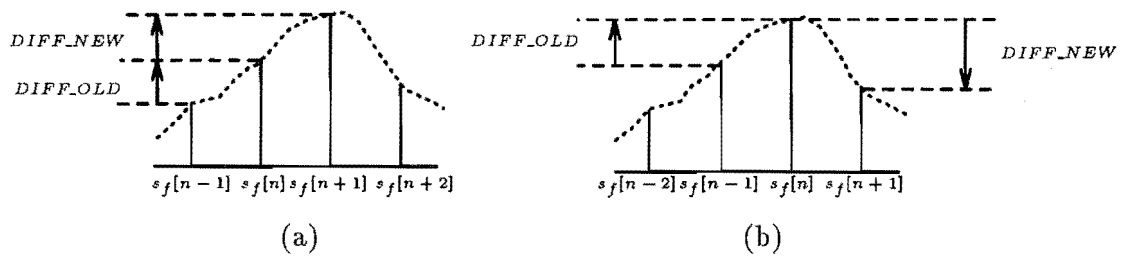


Figure 4.5. The extremum detection method: in (a) no extrema is identified at $s_f[n]$, the values of $DIFF_OLD$ and $DIFF_NEW$ are the same sign, in (b) an extrema is identified at $s_f[n]$ since the values of $DIFF_OLD$ and $DIFF_NEW$ differ in sign.

Both the amplitudes of the extrema and the times between the extrema are stored in a circular buffer in the program memory of the TMS32010. The circular buffer can store up to 64 amplitudes of extrema and 64 inter-peak times. Once all the extrema and inter-peak times from the 101 speech samples have been collected the next step is to calculate the pitch.

Figure 4.6 is the flow diagram of the peak detection method.

4.4.1.2 Pitch Extracting

The pitch estimation section of the pitch algorithm can itself be broken into two parts. The first part finds an estimate of the pitch and the second part checks this estimate. Before any pitch estimation is done a few preliminary operations must be performed (see fig 4.9). In addition, for the initial iteration of the pitch algorithm the previous pitch period, T'_0 , is set arbitrarily to be 10ms.

Next, the extremum with the maximum absolute amplitude is found. From this value the pre-set tolerance ε is calculated (see sec 4.4). For this pitch algorithm ε was set to be 20% of the amplitude of the maximum extremum (Chou, 1985).

The TMS32010 pitch algorithm estimates the pitch of voices that fall in the 100Hz to 625Hz range. It is actually the pitch period, rather than the pitch which is estimated by the algorithm. The pitch period is estimated from all the extrema in the circular buffer (see sec 4.4.1.1), not just the extrema extracted from one segment of 101 speech samples. The extrema in the circular buffer are updated for every 101 speech samples in such a way that the most recent extremum replaces the oldest extremum in the buffer. If the pitch period were only estimated from the extrema in the 101 speech samples, the algorithm could only extract pitches as low as 300 Hz. However the pitch period is estimated from the extrema in the circular buffer, which means that the pitch can effectively be checked over more than 10.1 ms (the duration of one segment of speech). The lowest pitch value the pitch algorithm can extract is set to 100 Hz. If the estimate of the pitch period exceeds 10ms (which corresponds to a pitch of 100 Hz), the algorithm ceases to estimate the pitch period for that segment of speech. How this is done is discussed later in this section.

Since the maximum expected pitch was 625 Hz there was no need to investigate pitch periods less than 1.6ms. To implement this the pitch estimate begins at the first extremum further than 1.6ms from P_m . Therefore the first comparison between the amplitudes of the stored extrema is between P_m and P_j where P_m is the most recently stored extremum and P_j is the next extremum occurring MIN_NUM_PK extrema after the most recently stored extremum P_m , where MIN_NUM_PK is the number of extrema within 1.6ms of P_m (MIN_NUM_PK is 1 in fig 4.7). If

$$|P_m - P_j| > \varepsilon \quad (4.8)$$

then no estimate of the pitch period is found, j is incremented and the next extremum amplitude is compared with P_m (see fig 4.7). This process is repeated until j is incremented to be a value p , where p is the smallest integer such that (4.4) is true. The time lapse between P_p and P_m is the first estimate of the pitch period (see fig 4.7). The estimate of the pitch period can be calculated using (4.5).

The algorithm was designed to cease calculating the pitch period if the pitch estimation became less than 100Hz. If the number of peaks between the extrema pointer

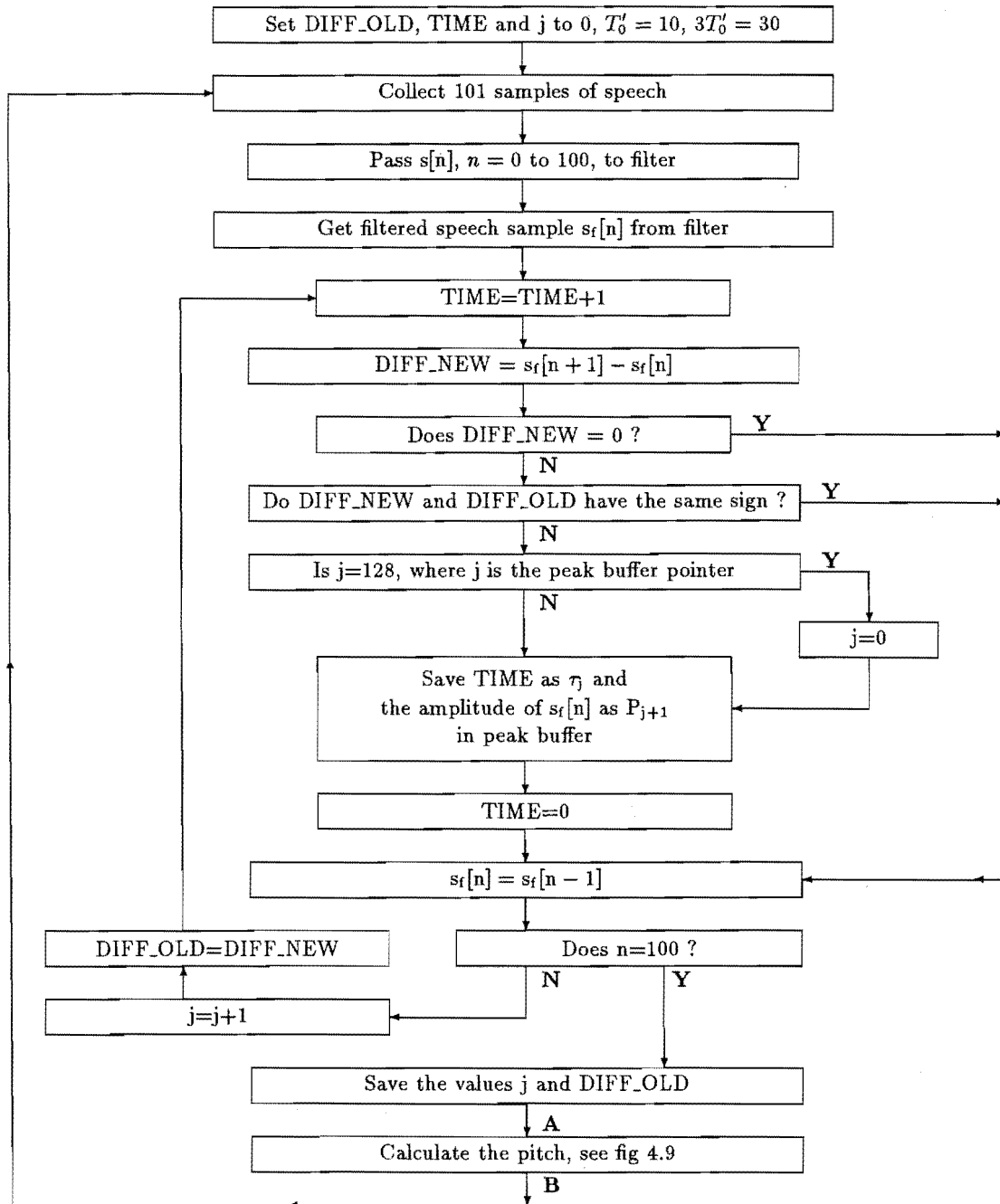


Figure 4.6. The flow diagram of the Extrema Detection subsection of the TMS32010 pitch algorithm.

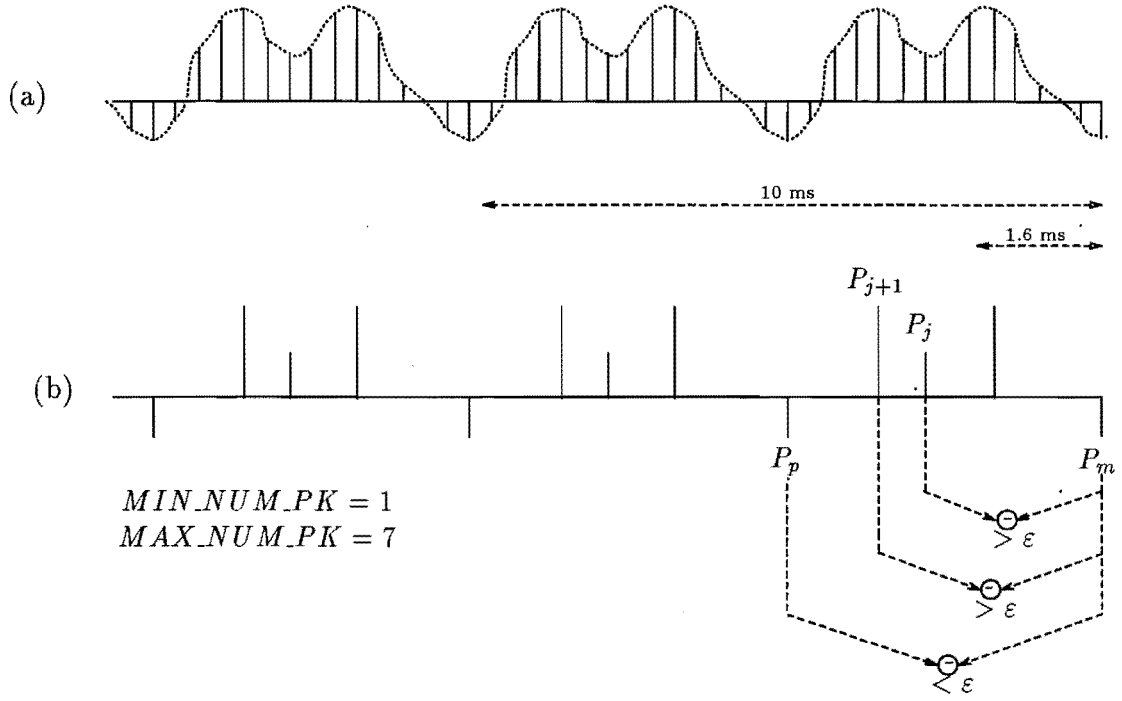


Figure 4.7. A pictorial representation on how to obtain an estimation of the pitch period using Brieseman's Pitch Algorithm, where (a) is the original waveform and (b) shows the extrema of that waveform.

j and P_m exceeds the value MAX_NUM_PK , the number of extrema within 10ms of P_m (MAX_NUM_PK is 7 in fig 4.7) then the time interval between P_m and P_j has become greater than 10ms. This means the estimation of the pitch will be less than 100 Hz. The pitch estimation procedure then halts for this segment of speech samples. The pitch period from the previous speech segment, T'_0 , is output to the IBM-PC and the whole pitch estimation process begins again with a new set of speech samples.

However if a value p is found (the values p, m and j are pointers to peaks in the peak buffer), such that (4.4) holds then an estimation of the pitch has been achieved. The next step in the algorithm is to check that this estimation holds for three times the previous pitch period, $3T'_0$; if it does a new estimate of the pitch has been found. The pointer m is incremented to $m + 1$, the next most recent amplitude, and p is incremented to $p + 1$, (see fig 4.8). If

$$|P_{m+b} - P_{p+b}| < \varepsilon, \quad (4.9)$$

where $b = 1$, is true and remains true for $b = 2$ to $3T'_0_NUM_PK - p$ ($3T'_0_NUM_PK$ is the number of extrema within $3T'_0$), then the pitch period value has been found (see fig 4.8). It is the sum of the inter-peak times between P_m and P_p (or P_{m+1} and P_{p+1} etc). The calculated value of the pitch period is then output to the IBM-PC.

If at any stage (4.9) becomes false, a new pitch estimate must be made. The algorithm returns to the pitch estimation part, comparing P_m and P_j , where P_m is the most recent extremum and $j = j + 1$. The process then proceeds as before.

Figure 4.9 is the flow diagram of the pitch extracting section of the TMS32010 pitch algorithm.

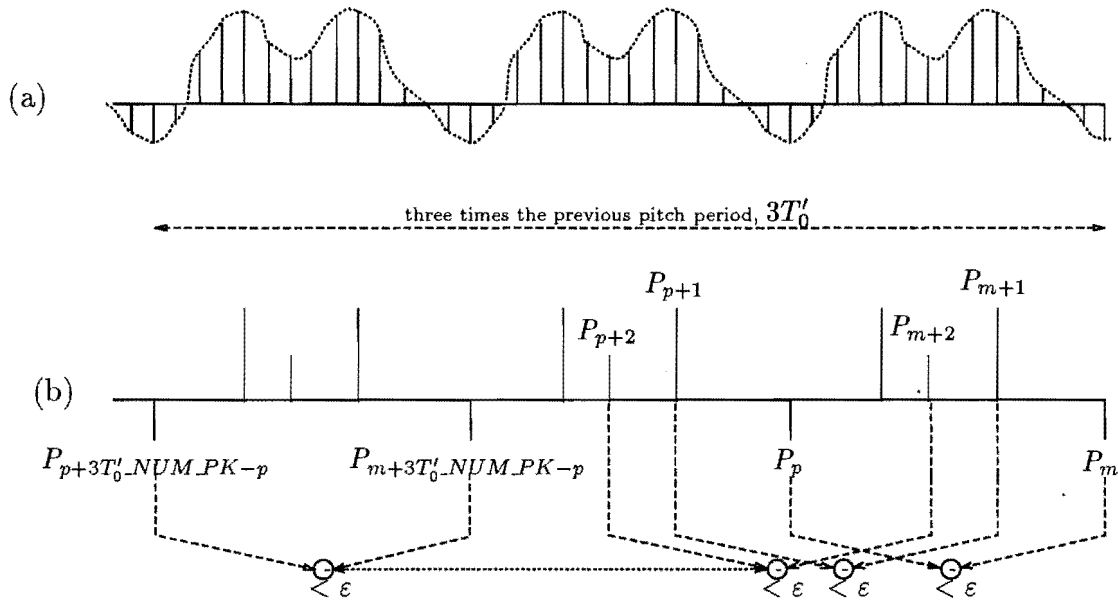


Figure 4.8. A pictorial representation how the pitch period estimate is checked in Brieseman's pitch algorithm, where (a) is the original waveform and (b) shows the extrema of that waveform.

4.4.2 The Modules Which Use The TMS32010 Pitch Algorithm

There are two modules in the CASTT which use the TMS32010 Pitch algorithm. These are the Voice Pitch Tracker and the Concurrent Loudness and Pitch modules. In both of these modules the TMS32010 Pitch algorithm is used to calculate the pitch values of the pitch contours.

4.5 SPECTRAL ANALYSIS

Different sounds ([m] and [i] for example) do not sound similar. This is because they have a different quality. Quality is the perceptual correlate of the frequency components of a sound, i.e. its spectral content. Spectrogram plots, for example the displays of the Spectrogram module in the CASTT, show the magnitude of the different frequency components as they vary with time.

In the Spectrogram module Fourier Analysis is used to convert the time-domain speech signal (as detected by the microphone) into the frequency domain. The code calculates the spectrum of a signal every 20ms, quantises the magnitude of each frequency component into a 2 bit number and also calculates the display co-ordinates of where the spectrum of each 20ms of speech is to be plotted on the screen. Before the actual TMS32010 spectrogram algorithm is presented, this section will briefly discuss Fourier Analysis and what considerations are necessary when performing Fourier Analysis on digital computers.

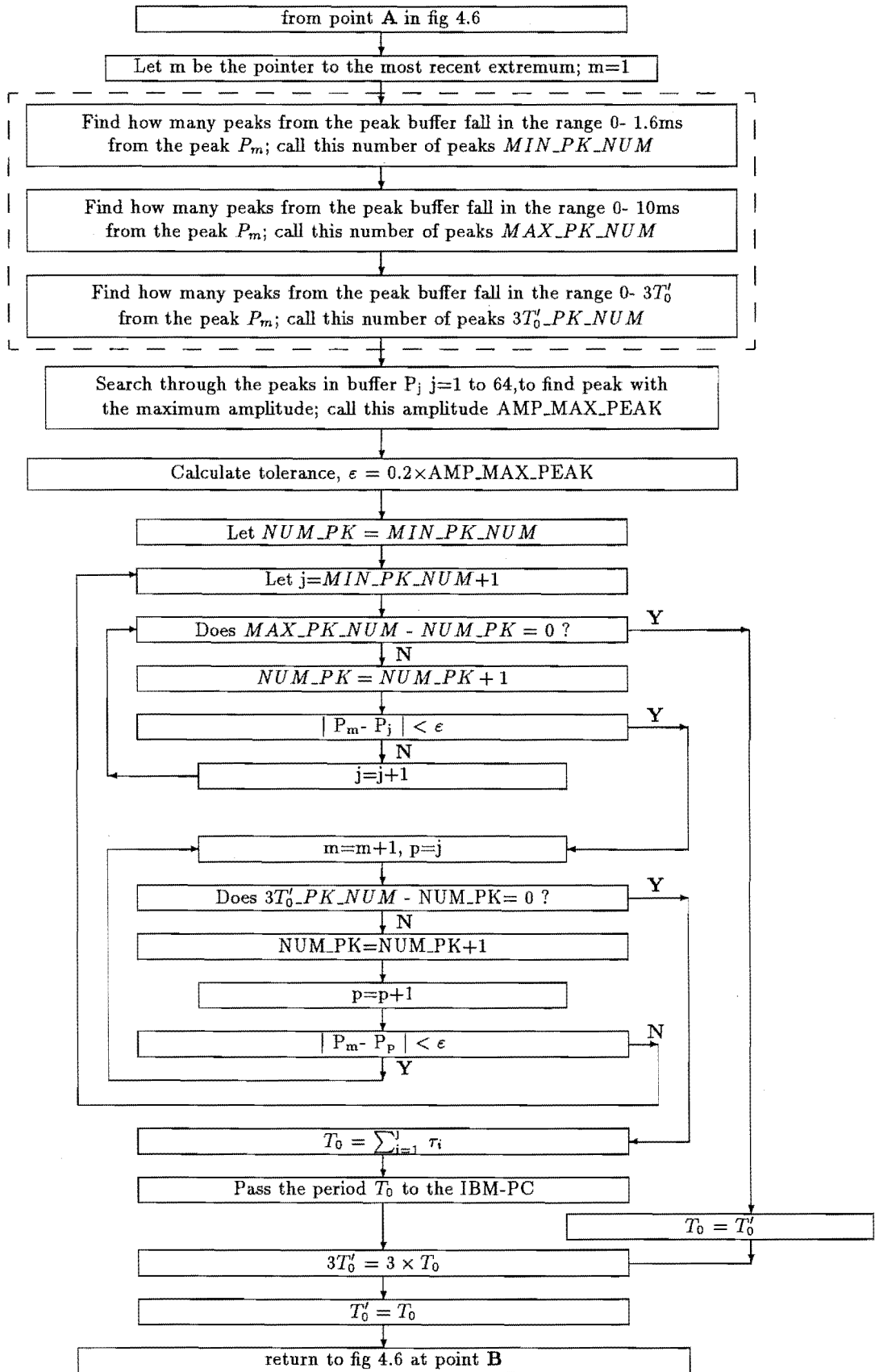


Figure 4.9. The flow diagram of the pitch extracting section of the TMS32010 pitch estimation algorithm.

4.5.1 Fourier Analysis

A periodic continuous signal, $s_p(t)$, can be resolved into an infinite sum of sine waves and cosine waves

$$s_p(t) = a_0 + \sum_{k=1}^{\infty} a_k \cos\left(\frac{2\pi kt}{T_0}\right) + b_k \sin\left(\frac{2\pi kt}{T_0}\right) \quad (4.10)$$

where a_k and b_k are the amplitudes of the cosine and sine waves, T_0 is the period and $\frac{k}{T_0}$ is the frequency of the k th harmonic of the fundamental frequency $\frac{1}{T_0}$. Equation (4.10) is known as the Fourier series of the signal $s_p(t)$. Using the identities

$$\cos(x) = \frac{1}{2}[e^{jx} + e^{-jx}] \quad (4.11)$$

and

$$\sin(x) = \frac{1}{2j}[e^{jx} - e^{-jx}] \quad (4.12)$$

the Fourier series can be rewritten in exponential form as

$$s_p(t) = \sum_{k=-\infty}^{\infty} s_k e^{j\frac{2\pi kt}{T_0}} \quad (4.13)$$

where the values s_k , for $k = 0, \pm 1, \dots, \pm \infty$, are the complex Fourier coefficients.

The complex coefficient s_k is calculated by the integral

$$s_k = \frac{1}{T_0} \int_{-\frac{T_0}{2}}^{\frac{T_0}{2}} s_p(t) e^{-j\frac{2\pi kt}{T_0}} dt \quad k = 0, \pm 1, \pm 2, \dots \quad (4.14)$$

For the Fourier series to hold, four conditions have to be satisfied. These are; (1) $s_p(t)$ is a single-valued function in the interval $-\frac{T_0}{2}$ to $\frac{T_0}{2}$, (2) $s_p(t)$ has at most a finite number of discontinuities in the interval $-\frac{T_0}{2}$ to $\frac{T_0}{2}$, (3) $s_p(t)$ has a finite number of maxima and minima in the interval $-\frac{T_0}{2}$ to $\frac{T_0}{2}$, and (4) $s_p(t)$ is absolutely integrable, that is

$$\int_{-\frac{T_0}{2}}^{\frac{T_0}{2}} |s_p(t)| dt < \infty \quad (4.15)$$

These four conditions are known as Dirichlet's conditions. They are satisfied by any bandlimited speech signal.

Equation (4.14) transforms continuous periodic time-domain signals into the discrete frequency-domain. Equation (4.13) performs the reciprocal function. To transform a non-periodic time-domain signal $s(t)$, into the frequency domain, to become the spectrum, $S(f)$, the continuous Fourier Transform is used:

$$S(f) = \int_{-\infty}^{\infty} s(t) e^{-j2\pi ft} dt \quad (4.16)$$

The inverse continuous Fourier Transform;

$$s(t) = \int_{-\infty}^{\infty} S(f) e^{j2\pi ft} df \quad (4.17)$$

transforms a continuous frequency-domain signal into the time domain.

The signal $s(t)$ is only transformable providing it satisfies the Dirichlet conditions, as mentioned above, where the interval $-\frac{T_0}{2}$ to $\frac{T_0}{2}$ is now $-\infty$ to ∞ . $S(f)$ is a complex function of frequency.

4.5.1.1 The Discrete Fourier Transform

The continuous Fourier Transform and Inverse continuous Fourier Transform are not suitable for digital signal processing. In digital signal processing both the time-domain signal and the spectrum must be discrete. A special form of Fourier analysis is needed for this situation. The Discrete Fourier Transform (DFT) transforms a discrete signal $s[n]$, comprising N samples into a discrete spectrum $S[k]$, comprising N samples:

$$S[k] = \sum_{n=0}^{N-1} s[n] e^{-j \frac{2\pi nk}{N}} \quad k = 0, 1, \dots, N-1 \quad (4.18)$$

The inverse discrete inverse transform performs the reciprocal function;

$$s[n] = \frac{1}{N} \sum_{k=0}^{N-1} S[k] e^{j \frac{2\pi nk}{N}} \quad (4.19)$$

The DFT cannot be thought of as merely a digitized version of the continuous Fourier Transform. If $s[n]$, $n = 0, 1, \dots, N-1$, in (4.18) and (4.19) is the digitized version of $s(t)$, in (4.16) and (4.17), then $S[k]$, in (4.18) and (4.19), will be the digitized version of $S(f)$, in (4.16) and (4.17) but only if no aliasing occurred in the sampling of $s(t)$ to get $s[n]$ and if there was no spectral leakage.

The spectral coefficients of a signal sampled in the time domain repeat with a cycle equal to $\frac{1}{T}$, where T is the sampling frequency. Therefore only frequencies in the range $-\frac{f_s}{2}$ to $\frac{f_s}{2}$, where f_s is the sampling frequency, are represented. Any frequency components greater than $\frac{f_s}{2}$ are therefore “folded over” and overlap the lower frequency components, causing aliasing. To avoid aliasing $s(t)$ must be band-limited and sampled at least the Nyquist frequency (see sec 4.1).

The following section discusses what spectral leakage is and how it can be reduced.

4.5.2 Windowing the Time Domain Signal

In speech processing, the processing is performed on sequential segments of the speech signal, rather than the entire signal. Thus Fourier analysis, either continuous or discrete, is performed on a signal of finite duration. The analysis reduces to constructing a Fourier Series of a signal with a fundamental frequency which is the reciprocal of the duration of the speech segment. If the signal contains spectral components which are not harmonic to this fundamental frequency the components calculated will not be accurately represented by the Fourier Series. This causes distortion which is called spectral leakage. The leakage can be better understood by considering that the Fourier Transform of a speech segment is actually the Fourier Transform of the entire speech signal $s(t)$ multiplied by a window rectangular function $w(t)$. In the frequency domain this results in the speech signal's spectrum $S(f)$ being convolved with $W(f)$, the Fourier Transform of the window function $w(t)$. The rectangular window function is defined as:

$$\begin{aligned} w(t) &= 1 \quad 0 < t < T_w \\ &= 0 \quad \text{otherwise} \end{aligned} \quad (4.20)$$

where T_w is the duration of the window. The Fourier Transform of $w(t)$ is:

$$\begin{aligned} W(f) &= \frac{\sin(2\pi f T_w/2)}{2\pi f} \\ &= \frac{T_w}{2} \text{sinc}(2\pi f T_w/2) \end{aligned} \quad (4.21)$$

The side lobes of the sinc function will distort the spectrum $S(f)$ if the duration of the window, T_w , is not a multiple of the period of $s(t)$.

Thus to completely reduce spectral leakage, the Fourier Analysis would have to be performed pitch synchronously. It is only feasible to calculate the Fast Fourier Transform on samples 2^x long, x being any positive integer. Pitch periods are hardly ever 2^x samples long, therefore it is usual practice to try and reduce the spectral leakage rather than eliminate it. This is done by windowing the speech segment with a window function which has lower side lobes than the rectangular function. The smaller the sidelobes, the less the spectral distortion. A commonly used window in speech processing is the Hamming window, which has smaller side lobes than the rectangular window (Owens, 1993, p47).

4.5.3 The TMS32010 Spectrogram Algorithm

The TMS32010 spectrogram program calculates a spectrum every 20ms from input speech using Fast Fourier Transform (FFT) analysis. The FFT is a computationally efficient algorithm for the calculation of DFT's (Brigham, 1974, p148). The TMS32010 spectrogram program can be split into two parts. Part one calculates the magnitude of the frequency components from the time-domain speech signals; part two assigns coordinates to each frequency component in the spectrum and assigns the colour of the pixel which represents the frequency component. The colour of the pixel is related to the energy of the frequency component.

4.5.3.1 Calculating The Magnitude Spectrum

The TMS32010 code calculates a 256 point FFT. 257 speech samples, $s[n]$, are collected and pre-emphasized using

$$s_{pe}[n] = s[n] - 0.95s[n-1] \quad (4.22)$$

where $s_{pe}[n]$, $n = 1$ to 256, is pre-emphasized speech. Pre-emphasis is a standard signal processing technique, which lifts the higher frequencies in the speech spectrum. There is an overall -6 dB/octave trend in the spectrum of voiced speech radiated from the lips (the reasons why are discussed in sec 8.3.2.1). Pre-emphasis, therefore, compensates for this roll-off of higher frequencies in voiced speech, the above operation in (4.22) lifting the spectrum by +6 dB/octave.

In unvoiced speech there is in fact an overall +6 dB/octave trend in the spectrum of speech radiated from the lips. There is no need to provide pre-emphasis for unvoiced speech, however for simplicity all the speech is pre-emphasised when it is processed in the Spectrogram module. The pre-emphasised speech is then windowed by a 256 sample Hamming window. This is to reduce the spectral leakage when the DFT is performed (see sec 4.5.2).

The DFT of the pre-emphasised windowed speech signals is performed using the Cooley-Tukey Radix Four FFT algorithm (Burrus and Parks, 1985). This transforms the discrete time-domain signal into a discrete spectrum in the frequency domain. The algorithm on the TMS32010 calculates a 256 point FFT in 10.6 ms (Elder *et al.*, 1987).

The Cooley-Tukey Radix Four FFT algorithm is a standard FFT algorithm, no modifications being made to it in the TMS32010 spectrogram program, so it will not be discussed in detail in this section. Interested readers can find more about the algorithm in Burrus and Parks' (1985) book "DFT/FFT and Convolution Algorithms".

For each segment of speech the Cooley Tukey Radix Four FFT algorithm calculates the real component, $\Re(S[k])$ and imaginary component, $\Im(S[k])$ of each of the complex frequency components $S[k]$, for $k = 0, 1, \dots, N-1$, in the spectrum. The displays in the Spectrogram and Sustained Phonation modules in the CASTT are in fact the spectra of the magnitude of the complex frequency components. The magnitudes of the spectral coefficients, $S(k)$, are computed by the approximation

$$|S[k]| \approx 0.96 \times \text{MAX}(|\Re(S[k])|, |\Im(S[k])|) + 0.398 \times \text{MIN}(|\Re(S[k])|, |\Im(S[k])|) \quad (4.23)$$

where $\text{MAX}(x,y)$ is the maximum value of x or y , and $\text{MIN}(x,y)$ is the minimum value of x or y . This approximation to the magnitude is accurate to within four percent of the actual value and is similar to the approximation used by Southard (1983). The magnitude is calculated using this approximation rather than through calculating the square root of the sum of the real and imaginary components because the approximation takes a shorter time to compute. There is insufficient time to calculate the magnitude from the square root and plot the spectrogram in real-time. Recall the TMS32010 has no in-built square root function.

4.5.3.2 Assigning Colour and Display Coordinates To The Data Points Of The Spectrogram

The TMS32010 spectrogram code not only calculates the spectrum, but it calculates the display co-ordinates of each point in the spectrum as well. This is because the 8088 processor of the IBM-PC is unable to calculate the display coordinates and plot them on a screen in real-time.

The CGA card supports four colours on the screen at any one time. This means the level of quantisation of the magnitude of the frequency components is limited to four levels. Hence they are quantised to a 2 bit number. Each of the numbers specifies a colour, which in turn represents a frequency component magnitude range. Only 128 frequency points on the spectrum need to be calculated for each 256 point FFT. Speech is a real time-domain signal. This means the resulting magnitude spectrum is even about $f = 0$. The portion of magnitude spectrum made from the negative frequency components is the mirror image of the portion of the magnitude spectrum made from the positive frequency components (Bracewell, 1978). This means the entire magnitude spectrum can be obtained from just calculating the portion of the spectrum in the positive frequency-domain (or conversely from the portion of the spectrum in the negative frequency domain).

Of the 128 frequency components produced in each spectrum only 64 magnitude values were passed to the IBM-PC. Every second component was discarded because a spectrogram of height 128 pixels would have taken up too much of the CGA 320 X 200 pixel screen. It would have enabled only one display plot to appear on the screen. The Spectrogram module, like the Loudness Monitor and Voice Pitch Tracker modules, has two display plots on the screen.

Once the pixel colours and positions of a spectrum have been passed to the IBM-PC the TMS32010 spectrogram program returns to the beginning of the algorithm to process the next 256 point FFT. The Program calculates the FFT with a 5.6ms overlap in the speech signal to account for the irregularities caused by windowing. This means that the new speech segment, for which the FFT is calculated, comprises the 57 most recent samples of the previous speech segment and the 200 new speech samples. Thus a new spectrum is calculated every 20 ms. The maximum length of speech for which the spectrogram can be calculated is 5.12 seconds.

The flow diagram for the TMS32010 algorithm is given in figure 4.10.

4.5.4 The Modules Which Used The TMS32010 Spectrogram Algorithm

The spectrographic plot calculated by the TMS32010 Spectrogram algorithm is displayed in two of the CASTT's modules. These are the Spectrogram Module and the Sustained Phonation Monitor. The difference between these modules is in the manner of presentation. The Sustained Phonation module has several animated cartoon features which the Spectrogram Module does not have (see sections 3.2.4 and 3.2.7).

4.6 ZERO-CROSSING ANALYSIS

The zero crossing rate for a segment of speech is the number of times the amplitude of successive speech samples changes sign. Sounds with high frequency components of high energy tend to have higher zero-crossing rates than sounds of high frequency components of low energy. Rabiner and Schafer (1978, p128) show that the zero-crossing rates for unvoiced sounds are generally greater than for voiced sounds. There is some overlap in the zero-crossing distributions however.

Fricative sounds are characterized by noise-like waveforms due to the air turbulence created in their production (Stevens, 1960). This means all fricative sounds have a large number of significant high frequency components, i.e. frequencies above 3 kHz. The zero-crossing rate of fricatives is generally greater than the rate for sounds that are not produced by turbulence. The energy for non-fricative voiced sounds tends to be concentrated below 3kHz (Rabiner and Schafer, 1978, p128). Ito and Robertson (1971) found zero-crossing measurements could be used in the detection of the fricative sounds [s], [ʃ], [f].

4.6.1 The TMS32010 Zero-Crossing Algorithm

The TMS32010 program calculated the zero-crossing rate for every 19.2ms of speech.

A zero-crossing for a discrete signal is said to have occurred if the sign between consecutive samples has changed. This can be detected by looking at the sign of the product of two consecutive speech samples. If the product is a positive number no

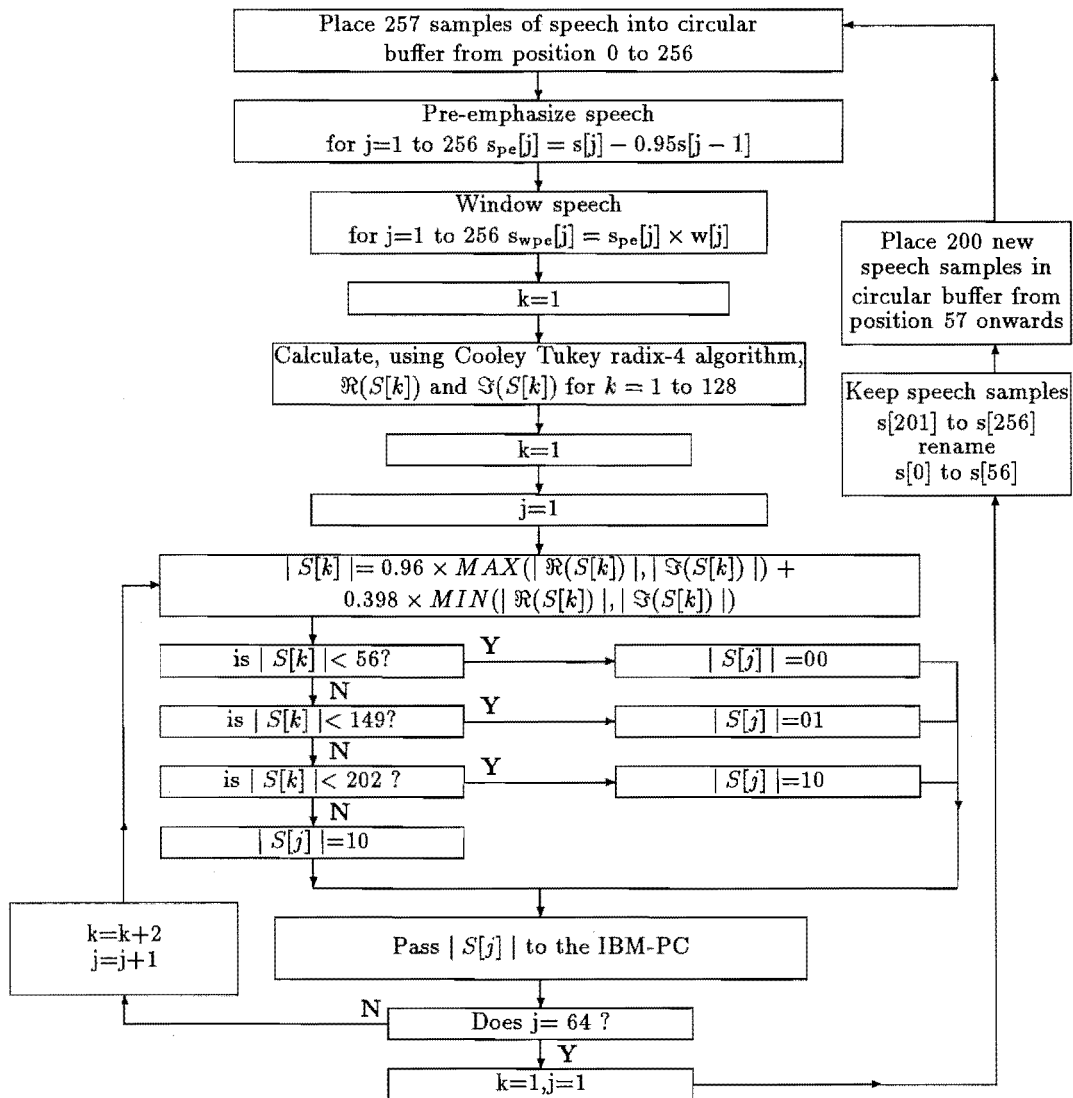


Figure 4.10. The TMS32010 spectrogram program algorithm which calculates a spectrogram in real-time.

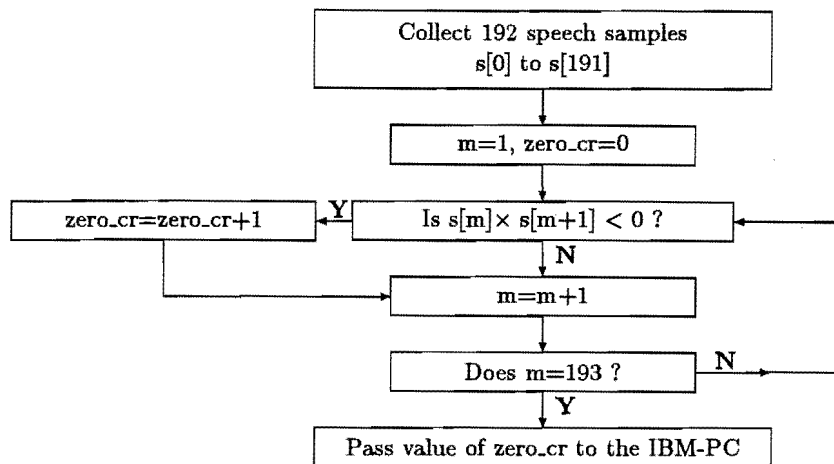


Figure 4.11. The TMS32010 zero-crossing algorithm which calculates the zero-crossing rate for 19.2 ms of speech and passes the result out to the IBM-PC.

zero-crossing has occurred. If the product is a negative number then a zero-crossing has occurred. Once the zero crossing rate for the 192 samples of speech has been calculated, the rate is passed out to the IBM-PC and the process begins again with the next 192 samples. The TMS32010 zero-crossing algorithm flow diagram is given in figure 4.11.

The Fricative Monitor in the CASTT utilises the zero-crossing technique. If the zero-crossing rate of a speech segment exceeds a pre-set threshold the sound is classified as a fricative and a horizontal bar appears on the screen. If the zero-crossing rate is less than the threshold the screen remains blank.

4.7 THE VOCAL TRACT RECONSTRUCTION

The final method of speech analysis used in the current modules of the CASTT is vocal tract reconstruction. This method of analysis is used in the Vocal Tract Shape module only. The Vocal Tract Shape module displays vocal tract shapes reconstructed from segments of speech. The reconstructed shapes are displayed against a mid-sagittal cross-section of the head, to give them a meaningful back-drop. The display of the Vocal Tract Shape module is constructed from the output of two algorithms. The first algorithm is, for the most part, executed on the TMS32010. It calculates the vocal tract area from segments of speech. The second algorithm is executed on the IBM-PC. It calculates the vocal tract shape and jaw position from the calculated vocal tract area. The Vocal Tract Shape module is the only module in the CASTT in which the signal processing for the final display is carried out in the IBM-PC, in addition to the TMS32010.

There are two cross-sections of heads on the screen. However the vocal tract shape can only be estimated for one speaker at a time. The vocal tract shape for the other head is frozen. For the head which is active, the vocal tract shape is updated every 40.4 ms.

4.7.1 The Vocal Tract Area Reconstruction Algorithm

The vocal tract area reconstruction algorithm is based on the Direct Estimation Method (Wakita, 1973). The algorithm is based on three main assumptions. The first assumption is that the vocal tract can be modelled as a series of concatenated lossless, rigid walled, acoustic tubes. The wave propagation down these tubes is linear and planar. The second assumption is that the vocal tract can be modelled as an all-pole filter. The final assumption is that, given certain conditions, the coefficients of the vocal tract filter are related to the reflection coefficients of boundaries between adjacent cross-sections in the acoustic tube module of the vocal tract. Provided the cross-sectional area of an acoustic tube at either the lips or glottis is known, the cross-sectional areas of the remaining concatenated acoustic tubes can be calculated from the reflection coefficients.

Markel and Gray (1976) have developed an efficient algorithm, based on the above assumptions, which makes it possible to calculate the cross-sectional areas of a vocal tract. The algorithm utilises the Durban and Levinson recursion algorithm. The algorithm used in the Vocal Tract Shape module in the CASTT is based on Markel and Gray's algorithm. To understand the algorithm completely one needs to have knowledge of the acoustic tube model of the vocal tract, the source filter model of speech and the linear prediction model of speech. These will be discussed in detail in sections 8.2 and 8.3.2.

The vocal tract area reconstruction algorithm is given in figure 4.12. From a speech segment of 101 samples the algorithm calculates the cross-sectional areas of the vocal tract A_i , $i = 0$ to 9.

4.7.2 Calculating the Vocal Tract Shape

The vocal tract area reconstruction algorithm calculates, from a segment of speech, the cross-sectional areas of 10 concatenated acoustic tubes. The reconstructed vocal tract shape can be formed from the 10 concatenated tubes. However in order to get the shape, manipulation of A_i for $i = 0$ to 9 is required.

In the Vocal Tract Shape module, the displays are updated each time a new set of cross-sectional areas are calculated. The entire display is not redrawn for each set of areas, however. The top of the head, the back of the vocal tract and the base of the throat (all indicated by the solid line in fig 4.13) remain unchanged, whatever the values of the cross-sectional areas. It is only the front part of the vocal tract (indicated by the dashed line in fig 4.13) and the position of the jaw (indicated by the dotted line in fig 4.13) that are updated. The outline of the top of the head and throat base are not updated because these parts do not move significantly in speech. The changing cross-section of the vocal tract is caused mainly by the movement of the jaw and the tongue. The back of the vocal tract is also assumed to remain rigid in speech. From the mid-sagittal cross-section of the vocal apparatus in fig 1.2, it can be seen that the back of the vocal tract between the larynx and the nasal cavity is against the spinal chord. The spinal chord restricts the movement of the back of the vocal tract. In addition the vocal tract bounded by the upper jaw bone, hard palate and soft palate can also be considered rigid.

Two separate algorithms are used to calculate the co-ordinates of the front of the vocal tract and the co-ordinates of the jaw. These two algorithms will now be outlined.

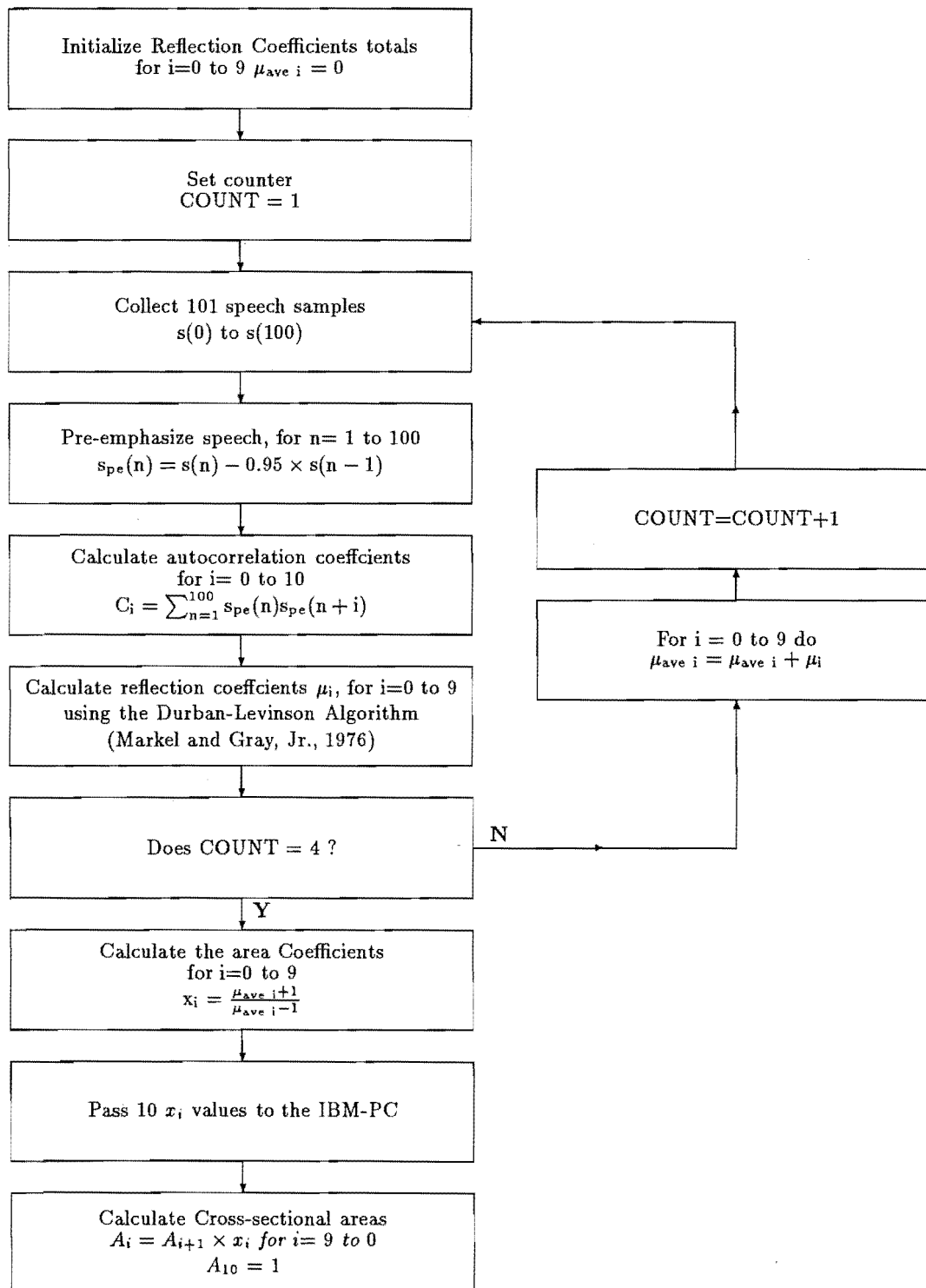


Figure 4.12. The algorithm which calculates the cross-sectional areas A_i , for $i = 0$ to 9, from which the vocal tract shape is estimated.

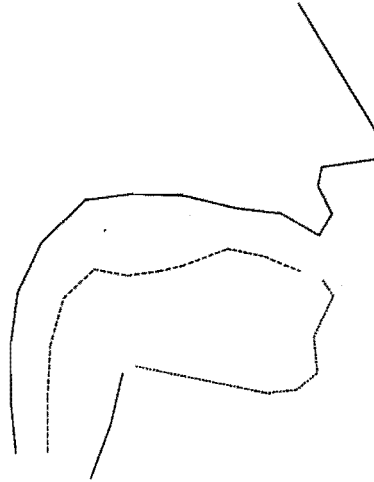


Figure 4.13. The outline of the mid-sagittal cross-section of the head. The solid line indicates the parts of the head that remain unchanged each time the vocal tract shape is updated. The dotted and dashed lines indicate the parts of the display that are continually updated.

| | lips | | | | | | | | | | glottis |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---------|
| | $i=0$ | $i=1$ | $i=2$ | $i=3$ | $i=4$ | $i=5$ | $i=6$ | $i=7$ | $i=8$ | $i=9$ | $i=10$ |
| $x_b(i)$ | 125 | 110 | 94 | 75 | 55 | 37 | 21 | 13 | 10 | 10 | 12 |
| $y_b(i)$ | 98 | 90 | 87 | 84 | 83 | 90 | 102 | 120 | 140 | 160 | 180 |

Table 4.1. The co-ordinates of the back of the vocal tract, $(x_b(i), y_b(i))$, where $i = 0$ to 10.

4.7.2.1 Calculating The Front Of The Vocal Tract Co-ordinates

The back of the vocal tract from the larynx through to the upper lip is approximated by 10 interconnecting chords of equal length. Associated with these 10 chords are eleven sets of co-ordinates $(x_b(i), y_b(i))$, where $i = 0$ to 10, which are given in table 4.1. The varying front of the vocal tract is also divided up into ten interconnecting chords. The eleven sets of co-ordinates defining these 10 chords, $(x_f(i), y_f(i))$, where $i = 0$ to 10, are obtained from projecting the diameters of the cross-sectional areas, d_i , $i = 0$ to 10, 90deg from the 10 chords which outline the back of the vocal tract (see fig 4.14).

In the acoustic tube model of the vocal tract the tubes all have the same length, it is only the diameters of the tubes that vary. We are only interested in reconstructing the shape of the vocal tract, not its actual dimensions. Therefore values of the diameters, d_i , $i=0$ to 10, are said to be the square root of the values of the cross-sectional areas of the acoustic tubes A_i , $i=0$ to 10. The co-ordinates of the front of the glottis, $(x_f(10), y_f(10))$ are set to be $(x_b(10)+10, y_b(10))$. One of the assumptions in the Direct Estimation method is that the cross-sectional area of the acoustic tube associated with the glottis is known (see sec 8.3.2). Therefore the value of the diameter of the acoustic tube associated with the glottis, d_{10} , remains the same regardless of what sound is being made.

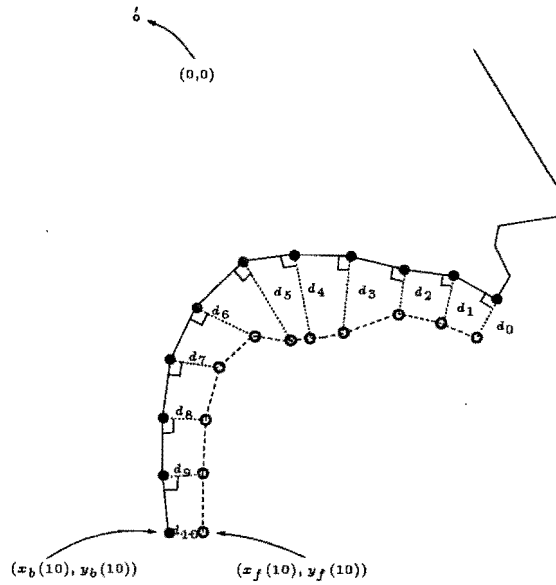


Figure 4.14. A graphical illustration of how the front of the vocal tract (indicated by the dashed line) is calculated from the back of the vocal tract (indicated by the solid line) and the diameters of the cross-sectional areas (indicated by the dotted line). The co-ordinates of the back of the vocal tract given in table 4.1, are indicated by the solid black dots. The co-ordinates of the front of the vocal tract, are indicated by the do-nut dots; the origin of the plot, $(0,0)$, is indicated on the diagram.

| | $i=0$ | $i=1$ | $i=2$ | $i=3$ | $i=4$ | $i=5$ | $i=6$ |
|-----------|-------|-------|-------|-------|-------|-------|-------|
| $jawx(i)$ | 60 | 110 | 120 | 127 | 124 | 130 | 125 |
| $jawy(i)$ | 136 | 141 | 138 | 131 | 116 | 101 | 98 |

Table 4.2. The co-ordinates of shape of the jaw, $(jawx(i), jawy(i))$, where $i = 0$ to 6.

4.7.2.2 Calculating The Jaw Co-ordinates

The lower jaw bone of a human (the mandible) is a u-shaped rigid structure. It pivots from two points, the Temporomandibular joints, which are very near the ears (Kaplan, 1971, p379). The Vocal Tract Shape module does not attempt to mimic exactly how the human jaw is pivoted in speech; it merely tries to get the position of the jaw correct.

The angle, θ_{jaw} , by which the jaw is rotated about its pivot is illustrated in fig 4.15 (a). It is calculated by:

$$\theta_{jaw} = \tan^{-1} \left(\frac{y_b(0) - y_f(0)}{120} \right) \quad (4.24)$$

where $y_b(0)$ is the y co-ordinate of the top lip, $y_f(0)$ is the y co-ordinate of the bottom lip. If the mouth were closed then θ_{jaw} would obviously be 0.

The jaw, as mentioned earlier, is a rigid structure. In the Vocal Tract Shape module the shape of the jaw is defined by the co-ordinates $(jawx(i), jawy(i))$ which are given in table 4.2. Once θ_{jaw} is calculated, the next step is to tilt the jaw shape by this angle. This process is illustrated in fig 4.15(b). The jaw is tilted θ_{jaw} from the closed mouth

position, indicated by the dotted lines in fig 4.15(b), to its new position, indicated by the solid line in fig 4.15(b).

Once the position of the front of the vocal tract and the jaw position have been calculated they are then plotted. To get a complete mid-sagittal cross-section of the head the front part of the vocal tract must be connected to the jaw, (see the dotted line in fig 4.15 (c) marked "1") and the jaw must be connected to the throat base (see the dotted line in fig 4.15 (c) marked "2"). Once this is down the entire plot is finished and ready to be updated.

4.8 LISSAJOUS FIGURES - A POTENTIAL SPEECH ANALYSIS MODULE FOR THE CASTT

Ideas for speech modules for the CASTT are continually being sought and considered. One idea that appears to have some potential is a module which displays Lissajous plots of speech signals. These plots can be thought of as a locus of the vector of an analytic signal in which the speech signal is the real component and the speech signal phase-shifted by 90 deg is the imaginary component. The idea was first suggested by Pronovost (1963) . Using this principle Pronovost found he could obtain discrete and distinguishable patterns for the sounds [i],[ɜ],[a],[o],[u],[m],[f],[v],[s],[z],[l] and [r](Pronovost, 1963; Pronovost *et al.*, 1968).

Pronovost built a dedicated hardware unit, called a Voice Visualizer, for the sole task of displaying Lissajous figures. The phase-shifting was achieved by an all-pass phase-splitting network, using the solid state technology available at the time. The speech signals were passed through an automatic gain control so the vocal intensity of the speech would not affect the size of the sound plots. The plots were all displayed on a cathode ray tube (Pronovost *et al.*, 1968).

The Voice Visualizer did not enjoy widespread popularity or use. However the usefulness of the Lissajous figure plots may have been prematurely dismissed or overlooked. The Voice Visualizer only had one speech analysis function, that is plotting Lissajous figures. This limited its applications and usefulness. The Voice Visualizer was an Articulation Corrector in which the display was a current value plot, (the display had no reference to time, see sec 2.3). This meant it could only be used for practising the pronunciation of isolated sounds. It could not be used for practising the sounds in words, sentences or spontaneous speech. The state-of-the-art technology used to build the system meant it would have been very costly for such a limited task.

Nowadays microprocessing technology makes it possible to have a computer-based speech therapy aid with many speech analysis functions. There is no single method of visually displaying speech characteristics which will be the most appropriate for the detection and correction of all speech disorders. Many different types of speech modules are needed in a speech aid. It was mentioned in sec 2.3.4 that for Articulation Correctors there is a need for both time-plot displays (those which show a record of the variations with speech features with time) and current-value-plots. For this reason there could be a place in a speech aid for a module which displays Lissajous figures of speech signals. The shapes of the Lissajous figures are distinctive for several different sounds. Young children may find creating the plots quite exciting; because of this I decided to build a software version of the Voice Visualizer.

The increased processing power of the digital signal processing chips meant that it

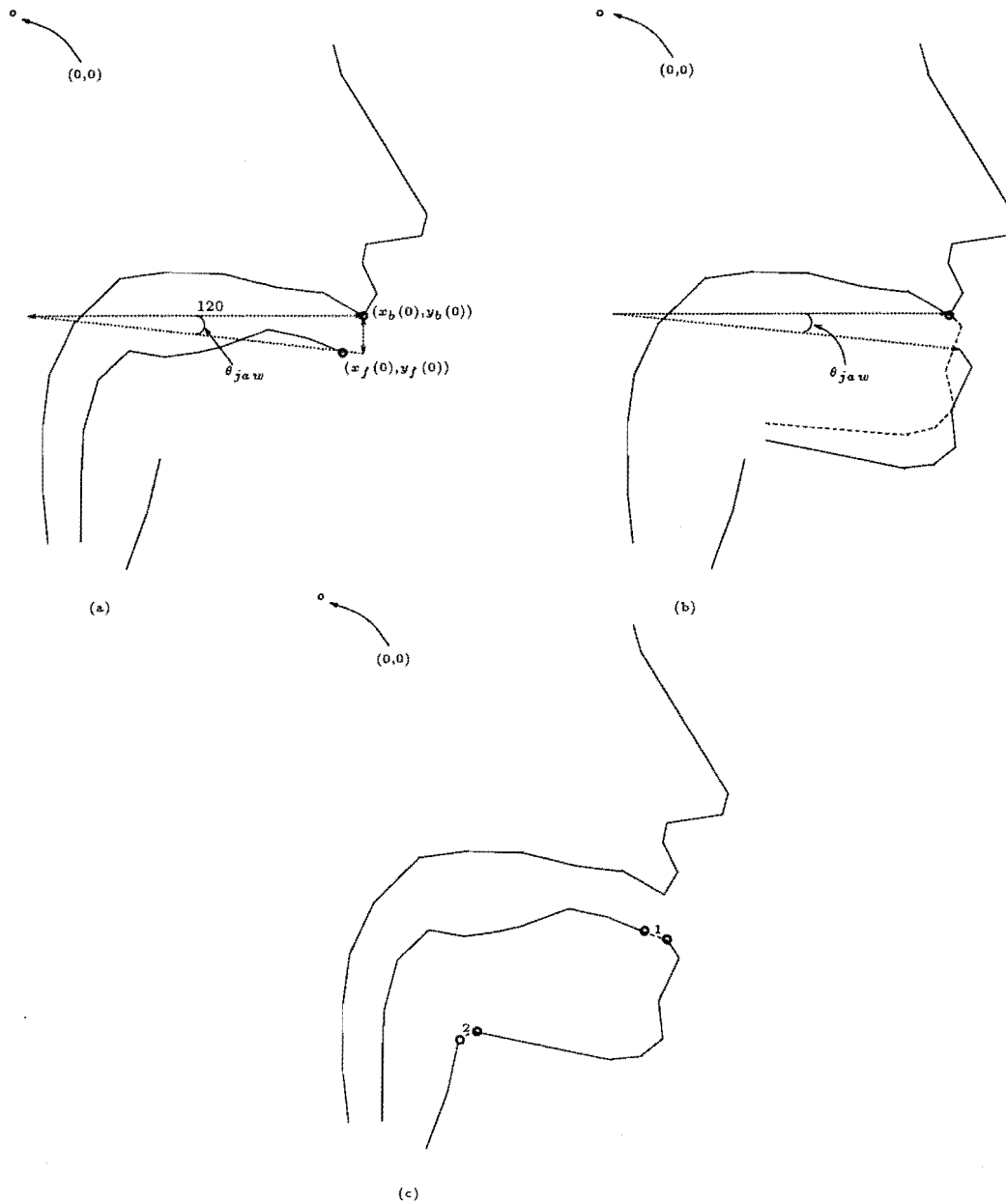


Figure 4.15. A graphical illustration of how the jaw position in the mid-sagittal cross-section of the head is calculated (a) firstly the angle the jaw is rotated about a pivot, θ_{jaw} , is calculated, (b) next the shape of the jaw is rotated by θ_{jaw} about this pivot (for simplicity the outline of the front of the vocal tract has been removed in this diagram), (c) finally the jaw is connected to the front of the vocal tract and to the base of the throat (in all three diagrams the origin $(0,0)$ is indicated).

was feasible to develop a real-time phase-shifting system in software. However before a real time Lissajous figure module could be developed an algorithm for phase-shifting the speech signal was needed. A phase-shifting algorithm was developed and tested in SIGPROC, the signal processing package developed for the departmental VAX mini-computer. The phase-shifting algorithm will now be described.

4.8.1 Phase Shifting In the Frequency Domain

To phase shift a signal by 90deg, 90deg is added to its positive frequency components and -90deg is added to its negative components. This will be illustrated by way of a simple example. Let the time-domain signal $h(t)$ be:

$$h(t) = \cos(2\pi f_c t) \quad (4.25)$$

Transforming this into the frequency domain, via the Fourier transform we get:

$$H(f) = \frac{1}{2}(\delta(f - f_c) + \delta(f + f_c)) \quad (4.26)$$

To phase shift $h(t)$ by 90deg is the same as adding 90deg to the positive frequency components of $H(f)$, (i.e. $\frac{1}{2}\delta(f - f_c)$), and -90deg to the negative frequency components (i.e. $\frac{1}{2}\delta(f + f_c)$). This is the same as multiplying the positive and negative frequency terms by j and $-j$ respectively. Thus (4.26) becomes

$$\begin{aligned} H_{90\text{deg}}(f) &= \frac{1}{2}(j\delta(f - f_c) - j\delta(f + f_c)) \\ &= \frac{-1}{2j}(\delta(f - f_c) - \delta(f + f_c)) \end{aligned} \quad (4.27)$$

Now taking the inverse Fourier transform of (4.27) we get:

$$h_{90\text{deg}}(t) = -\sin(2\pi f_c t) \quad (4.28)$$

Thus it can be seen that the original cosine signal has been phase shifted by 90deg and has become a sine wave (as expected).

4.8.2 The Lissajous Figure Algorithm

An algorithm giving Lissajous figures of speech signals was programmed on the departmental VAX using SIGPROC. An automatic gain control was used on the incoming signals to ensure all the plots were the same amplitude, regardless of the vocal intensity of the spoken utterance. The phase-shifting was performed by the method described above (see sec 4.8.1). Further processing was required in order to make the Lissajous plots like those obtained in the Voice Visualizer. The plots obtained from SIGPROC were angular, while in the Voice Visualizer they were smooth circular figures. The Voice Visualizer patterns were made from continuous signals whereas the Lissajous figures in SIGPROC were from discrete signals. To smooth the SIGPROC Lissajous figures the sampling rate of the digitized speech was increased to 80kHz by digitally interpolating between existing samples. Figure 4.16 shows the difference between the smoothed and unsmoothed Lissajous figures for the same speech segment of the sound [s].

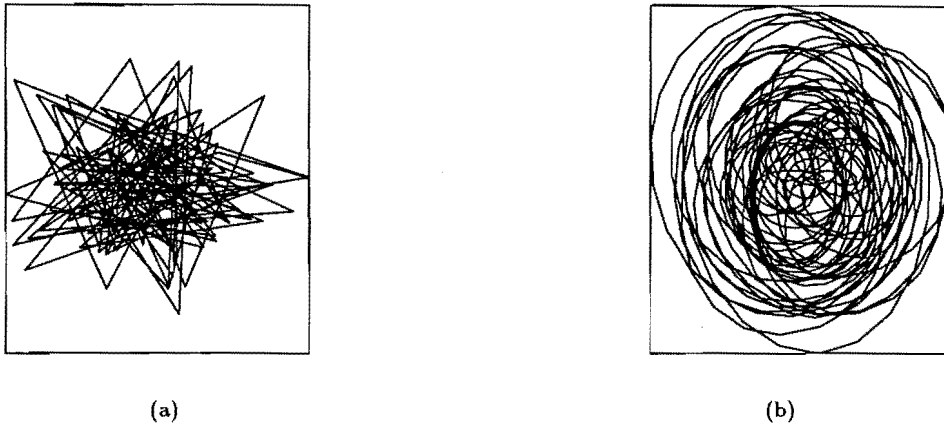


Figure 4.16. The Lissajous figures of the sound [s] plotted from (a) the speech signal and the speech signal phase shifted by 90 deg and (b) the Lissajous figure plotted from the high (80kHz) sampling rate versions of the speech signal and the phase-shifted speech signal.

Before adding a Lissajous figure module to the CASTT, it was decided it was necessary to test the potential of the display for speech remediation. The tests the Lissajous figures underwent, along with the rest of the displays in the CASTT, are presented and discussed in Chapter 6, as are the results of the tests. However before the evaluations of the CASTT's visual displays are presented, the evaluations of the CASTT by Speech Therapists will be given. The therapists' evaluations, and the consequences thereof will be presented in the following chapter, Chapter 5.

CHAPTER 5

EVALUATION OF CASTT BY SPEECH THERAPISTS

An important part of the CASTT's development has been its evaluation by speech therapists. The CASTT is not a visual speech aid developed in isolation in the Department of Electrical and Electronic Engineering at the University of Canterbury. It has been evaluated by 15 speech therapists in the Canterbury region. The evaluations have been on-going since I took over the project.

The ability to be able to display features of speech in a visual speech aid, alone, does not guarantee that the aid will be of any use in speech therapy. The aid must be easy for the therapist and the client to use, it must impart useful information and it must fit the requirements of speech therapy. These aspects cannot be guessed by the engineer; they can only be obtained from the therapists and clients themselves. Therefore in order for the CASTT to be an effective aid it was important to develop it with continual feed-back from speech therapists. Whilst this notion may seem like common sense, very few visual speech aids seem to have been built in this manner. Of the aids reviewed in sec 2.2 and 2.3 only the developers of the BBN aid (Boothroyd *et al.*, 1975), John Hopkins aid (Ferguson *et al.*, 1988) and VSA aid (Arends *et al.*, 1991) acknowledged the importance of on-going interaction with the speech therapy community during the development of their respective aids. From the literature it is clear that the ISTRa aid and the IBM-Speech Viewer were also developed closely with speech therapists; however no mention was made of the importance of this. (ISTRa: Osberger *et al.* (1978), Osberger *et al.* (1981), Osberger *et al.* (1982), Kewley Port *et al.* (1987a), Kewley Port *et al.* (1987b), Watson *et al.* (1989) and Kewley Port *et al.* (1991)) (IBM Speech Viewer: Denoix (1984), El Beze (1986), Adams *et al.* (1989) and IBM (1988)). Most of the aids reviewed in sec 2.2 and 2.3 do not appear to have been developed interactively with Speech Therapists. They were only evaluated by the therapists once there was a final product (see for example Pickett and Constam (1968), Thomas (1968) and Pardo (1982)).

The sentiment behind the development of the CASTT was very similar to that of the BBN aid (see sec 2.3). Boothroyd *et al.* (1975) said of the developers of the BBN aid: "it was their belief that an effective way to develop a worthwhile system would be to place a working, but not necessarily optimised, device into the hands of the [speech therapists] and to continue its development through a process of interaction."

The purpose of getting the speech therapists to evaluate the CASTT was twofold. Firstly, it was hoped the evaluations would provide the direction for the development of a better system. Secondly the evaluations were intended to indicate the potential of the CASTT as a speech aid. Information was obtained from the evaluations on both these points. Many changes and improvements were made to the CASTT as

a result of the therapists' comments (see sections 5.1.1.1 and 5.1.2.1). Qualitative evidence was obtained that the speech of some clients did improve when they used the CASTT in speech therapy (see sec 5.2.1). Some of the evidence, however, was questionable. One therapist cited two examples of successfully using the CASTT to remediate speech errors it could not possibly provide useful information on. This meant it was questionable whether these improvements could be attributed to the CASTT. Not only that, it was, therefore, questionable whether any of the improvements observed by any of the therapists could be attributed to the CASTT since it had not been established exactly what information one could expect to acquire from the displays of the CASTT. Therefore a third consequence of the therapists' evaluation of the CASTT was that we saw the need to develop a test to assess the visual displays of the CASTT to establish what sort of information we should expect to get from the aid.

The BBN, VSA, ISTR, IBM-SpeechViewer and John Hopkins aids (see sec 2.3) were all products of research teams comprising members with backgrounds in engineering and others with backgrounds in speech therapy. This combination ensured that the visual speech aids fitted the requirements of speech therapy, rather than that the speech therapy for a client had to be based around the capabilities of the aid. Many speech aids have not fulfilled this requirement. Lippmann (1982) commented "Training procedures have often been designed primarily on the basis of the capability of the aid, rather than on the basis of a theoretical model of speech acquisition or by considering the capabilities of the student and the existing speech training program."

The CASTT was not developed under such ideal conditions as the BBN, VSA, ISTR, IBM-SpeechViewer and John Hopkins aids. Whilst the CASTT was evaluated by many speech therapists, the only people who conducted research on the CASTT were Electrical Engineers, mainly myself (see sec 3.1). The therapists who evaluated the CASTT had no professional commitment to the project. They evaluated the CASTT only out of interest. This meant two things. Firstly there was no pool of speech therapists readily available to evaluate the CASTT. Secondly I had to be very careful not to lose the goodwill of the therapists.

To optimise the feedback from the therapists, the CASTT was evaluated in a series of bursts by several speech therapists. The therapists who initially partook in the evaluations were contacted by me and asked if they would like to become involved in evaluating the CASTT. As I became more known to the Canterbury Speech Therapy Community several of them approached me to see if they could use the CASTT.

The first type of evaluation of the CASTT involved the aid being placed into a speech therapist's clinic for a 2-3 week duration. This will henceforth be called a short-term evaluation. In these evaluations twelve different speech therapists took part. In the second type of evaluation of the CASTT, the therapist had the CASTT in their clinic for at least two months. This will henceforth be called a long-term evaluation. Four speech therapists were involved in these. The evaluations of the CASTT will now be discussed.

5.1 THE SHORT-TERM EVALUATIONS OF THE CASTT

The first series of evaluations to be performed on the CASTT were short-term evaluations. There were two separate periods when the CASTT underwent short-term evaluations. These will be outlined separately.

5.1.1 The First Assessment Period Of The CASTT

The first version of the CASTT which was assessed by speech therapists comprised five speech analysis modules. The modules were the Voice Pitch Tracker, the Loudness Monitor, the Fricative Monitor, the Vocal Tract Shape Module and the Spectrogram Module. The versions of all these modules, except the Fricative Monitor module, were earlier and more rudimentary ones than those presented in chapter 3. The Vocal Tract Shape Module, for example, only displayed one mid-sagittal cross-section of a head on the screen. At the time of the first assessment period I had made no contribution to the software of the CASTT. All five modules were written by previous participants in the CASTT project (see sec 4.2). I had, however, by the time of the first assessment period become the sole person involved in the CASTT project.

The first assessment period took place in mid-1988 with six speech therapists. Four had their clinics in a hospital environment. The clients in these clinics were all adults. The other two speech therapists in the first assessment period were in clinics associated with schools. One clinic was at a specialist school for the hearing-impaired. The other was at a school in a special section for the physically disabled. The clients at these last two clinics were all children.

The Voice Pitch Tracker, Loudness Monitor and Fricative Monitor modules had been suggested by various speech therapists. However, until the first assessment period no therapist had used any of the modules in the CASTT in their clinic with clients. The intention of this evaluation was, essentially, to see how the speech therapists reacted to the modules and to see what improvements could be suggested.

The CASTT, when placed in the speech therapy clinic, consisted of an IBM-PC XT with the TMSboard (see sec 4.2), a microphone, floppy disks and a help manual. In the version of the CASTT used by the therapists in the first assessment period the floppy disk booted up the IBM-PC XT and contained the executable files of the speech modules. To run the speech modules it was necessary to have knowledge of a few IBM-PC DOS commands and know how to load floppy disks. Each therapist was given an hour-long tutorial session about the CASTT. In that session they were shown how to run the speech modules and what each module did.

To formalise the therapists' response about the CASTT they were asked to fill out a questionnaire on the CASTT. A copy of the questionnaire is given in appendix A.1.

5.1.1.1 The Consequences Of The First Assessment Period

After the first assessment period many changes to the CASTT were made as a result of the suggestions by the speech therapists in the questionnaires. In fact this was the time when most of the changes were made to the actual modules of the CASTT. This can be seen in fig 5.1, which illustrates the evolution of the CASTT, from the five rudimentary modules to the seven modules described in chapter 3, relative to the speech therapists' evaluations of the CASTT.

The module which was given the most extensive overhaul, due to the feedback from the therapists in the first assessment period, was the Voice Pitch Tracker. Several changes were made. A criticism voiced by several therapists of the Voice Pitch Tracker was that it lacked sensitivity to the pitch variations in high pitched voices (mainly those of women and children). This, it transpired, was due to an artefact in the TMS32010 Pitch algorithm at the time. In the current version of the Voice Pitch Tracker module

| | |
|--|--|
| Prior to author's involvement in CASTT project | Rudimentary versions of : Fricative Monitor; Voice Pitch Tracker; Loudness Monitor; Spectrogram Module; Vocal Tract Shape module. |
| First assessment period of CASTT | Changes made to Voice Pitch Tracker: <ul style="list-style-type: none"> • made sensitive to high pitch variations; • Overlay command included; • Data Page option included; • made to only start plotting contour from beginning of speech; • new format of option menus. Changes made to Loudness Monitor: <ul style="list-style-type: none"> • Overlay Command included; • made to only start plotting contour from beginning of speech; • new format of option menus. Changes made to Vocal Tract Shape Module: <ul style="list-style-type: none"> • two mid-sagittal cross-sections of head; • Overlay Command included; • vocal tract shape assumes neutral position if silence detected for more than 0.404 secs. Addition of Concurrent Loudness and Pitch Module. Addition of CASTT software shell. |
| Second assessment period of CASTT | |
| | Addition of Sustained Phonation Monitor. |
| Long-term Evaluation of CASTT | The version of the CASTT as described in chapter 3. |

Figure 5.1. The evolution of the CASTT relative to the evaluations performed on the CASTT by the Speech Therapists.

the pitch contours are obtained from the estimation of the fundamental frequency of sequential segments of speech (see sec 4.4). In the version of the module at the time of the first assessment period, the pitch contours were not actually plots of the fundamental frequency. Rather they were the plots of the reciprocal of a value T' , T' being defined as:

$$T' = (148 - T) * 2 \quad (5.1)$$

where T is the estimated pitch period of sequential speech segments, calculated by the TMS32010 pitch algorithm outlined in figures 4.6 and 4.9. This manipulation of the pitch period estimate in (5.1) meant the resulting pitch contours were more sensitive to variations of low pitched voices (mainly men) than to the variations in the pitch of high pitch voices (mainly women and children). No reason was given for the presence of the manipulation of the pitch period estimate. However since the original developer of the TMS32010 Pitch algorithm was a man, it is possible the manipulation was done in order that the resulting pitch contours were sensitive to his pitch variations. There is a non-linear relationship between fundamental frequency and its perceptual correlate, pitch (Clark and Yallop, 1990, p210). As a result the frequency range in the variations of higher pitch voices are always greater than the frequency range in the variations of low pitched voices.

The pitch contours obtained directly from the estimate of the pitch period are more sensitive to the variations in higher pitched voices than the variations in lower pitched voices. Most of the speech therapists were women and the vast majority of the clients the CASTT was tested on were children. Therefore it seemed sensible to remove the manipulation from the TMS32010 Pitch Algorithm. The Pitch algorithm in the TMS32010 was altered to the one discussed in sec 4.4. Thus the resulting pitch contours became more sensitive to variations in high pitched voices than to variations in low pitched voices (see fig 5.1). This, of course, was only seen as a temporary measure. Ultimately two version of the Voice Pitch tracker are needed, one for high pitched voices and the other for low pitched voices.

The Overlay command and the Data Page option in the Voice Pitch Tracker module (see sec 3.2.1) were added to the module as a result of suggestions by speech therapists in the first assessment period. The Overlay command (see sec 3.2.2) was added to the Loudness Module, for the same reason.

The other module that was extensively modified, as a result of feedback from the speech therapists, was the Vocal Tract Shape module (see fig 5.1). At the time of the first assessment period the display consisted of only one mid-sagittal cross-section of the head. This was increased to two, as a result of a suggestion of a speech therapist. It should be noted that the vocal tract shape can only be updated on one mid-sagittal cross-section at a time. An option was introduced in the module which enabled the reconstructed vocal tract shape of the therapist to be overlayed on to the client's shape. This too was an idea from a speech therapist.

As a result of another suggestion from a therapist in the first assessment period one new module was added to the CASTT. This was the Concurrent Loudness and Pitch module (see fig 5.1).

Several other changes were made to the CASTT after the first assessment period. These were not due to suggestions of the therapists; however they came about from observing the therapists using the CASTT. It often took the clients (and even the

speech therapists) some time to speak once the modules had been placed in the active mode. This meant that in the Voice Pitch Tracker and Loudness Monitor modules the pitch or loudness contours were being calculated for that initial silence. The pitch and loudness contours have zero amplitude for silence. The length of the initial part of the contour of zero amplitude only indicated the duration of the silence between the time the module was placed in the active mode and the time when a sound was detected by the microphone. As a result the Loudness Monitor and Voice Pitch Tracker modules were altered so the plotting of the contours began once a sound was made, rather than as soon as the module was placed in the active mode (see sec 4.3.2). This feature was also added to the Concurrent Loudness and Pitch module (see sec 4.3.2) and a similar feature was added to the Vocal Tract Shape module. The continually updated vocal tract shape assumed a default vocal tract shape if silence was detected by the microphone for more than 0.404 seconds (see sec 4.3.2). It remained in this position until speech was detected. This feature was added because in the rudimentary version of the Vocal Tract Shape module the vocal tract shape was being continually updated regardless of whether a sound was spoken into the microphone or not. This could become very distracting.

The final changes made to the CASTT (see fig 5.1) were to make it completely menu driven and to unify the format of the menus. All the speech modules of the CASTT in the first assessment period had to be executed from DOS. This led to some confusion for the speech therapists. The commands and options within the modules were menu driven. A new option or command was invoked each time an appropriate key was pushed. However all the modules had to be run from the DOS command line. The name of the module had to be typed in, then the return key had to be pushed. For people not used to using a computer it was difficult for them to realize when they were to use DOS commands or when they only needed to push a key to invoke a command or option. To solve this problem all the modules in the CASTT were incorporated within a single package so that the modules could be run from within a software shell.

The format of the menu for the CASTT software shell, from which the modules can be run, was made to be the same as those for the option menus in the Fricative Monitor, Spectrogram and Vocal Tract Shape modules. The format of the option menu in the new Concurrent Loudness and Pitch module was also the same as for those modules. The format for the option menus in the rudimentary versions of the Voice Pitch Tracker and Loudness Monitor modules, however, differed from that for the rest of the modules. This reflected the fact that the CASTT modules had been developed by more than one person. To unify the user interface throughout the CASTT, the format of the option menus of the Voice Pitch Tracker and Loudness Monitor modules was changed to be the same as the other modules.

In the first assessment period no specific feedback was obtained on the usefulness of the modules. This was not surprising since the 2 -3 week duration of the evaluations was not long enough for the speech therapist to work out a training program using the CASTT. However there was a general feeling amongst the therapists that the CASTT had potential as an aid and that the development of the aid was worth pursuing.

5.1.2 The Second Assessment Period

After all the changes discussed in sec 5.1.1.1 had been made to the CASTT, it was once more placed into several of the speech therapy clinics. There were other changes suggested by the therapists in the first assessment period that could have been made to the CASTT (such as the inclusion of a Sustained Phonation Monitor (see sec 5.1.2.1)). However it was felt it was time to see how the therapists reacted to the numerous changes already made to the CASTT. The second assessment period took place in mid 1989. Seven speech therapists were involved. All of them had clinics associated with schools and all the clients were children. Only one of the therapists in the second assessment period was in the first as well. One other therapist in the second assessment period had experimented with the CASTT speech modules on her own, but she had not used them with her clients. No therapists associated with hospitals were involved in the second assessment period. This was because at the time of the assessment none of those therapists were available.

The purpose of this series of evaluations was (i) to see the effectiveness of the changes made to the CASTT, (ii) to see what other improvements might be suggested and (iii) to gain an idea of the CASTT's potential as a speech aid. The version of the CASTT that the therapists received in their clinics was that which was described in chapter 3, except that it did not yet include the Sustained Phonation Module. It contained six speech analysis modules, the Voice Pitch Tracker, the Vocal Intensity Monitor, the Concurrent Loudness and Pitch module, the Vocal Tract Shape module, the Fricative Monitor and the Spectrogram module. Each therapist was given an hour-long tutorial on the aid, as in the first assessment period. In the tutorial they were shown how to run the CASTT software package, what each speech module was designed to display and what analysis options were available in each module.

Once again, to formalise the speech therapists' responses to the CASTT, they were required to fill out a questionnaire about it. The questionnaires in the first and second assessment periods were different. A copy of the questions asked in the questionnaire of the second assessment period is listed in appendix A.2. The first questionnaire contained 29 questions, the second questionnaire contained 76. Due to the large amount of information obtained in the first assessment period questionnaire it was mistakenly thought that if more questions were asked more information would be gained. This was not the case. In fact it became apparent after the second assessment period that compiling a useful questionnaire was not a trivial task. Both questionnaires were compiled by myself, these two questionnaires being the first ever composed by me. This thesis is not going to discuss the problems associated with compiling a questionnaire nor will it discuss the merits and limitations of the questionnaires used in the two assessment periods. It is well beyond the scope of this thesis. Suffice it to say that there are many pitfalls associated with the composition of questionnaires.

5.1.2.1 The Consequences Of The Second Assessment Period

Unlike the first assessment period, there was not a large number of quickly implementable changes that were suggested in the second assessment period. Only one change was made to the CASTT after the second assessment period and that was the inclusion of a Sustained Phonation Monitor module (see fig 5.1). The inclusion of this module was in fact just as much a consequence of the first assessment period as it was

of the second. In both assessment periods, several therapists had commented that they thought the display of the Spectrogram module was too crude and coarse to be of any use. However they did say that the display was very pretty and captured the children's interest. One speech therapist in the first assessment period had suggested a new use for the Spectrogram module. She suggested it be used to promote sustained phonation. If one sound was sustained the spectrogram pattern remained fairly constant. If the speaker slipped into phonating another sound the change in pattern was noticeable.

Instead of renaming the Spectrogram module the Sustained Phonation Monitor, it was decided to develop a separate module. The Spectrogram module was not replaced because some of the other therapists in the two assessment periods had thought it was a good module. Whilst the plots of the Spectrogram and Sustained Phonation Monitor are calculated in the same manner, the display format for the two modules is quite different (compare fig 3.4 with fig 3.7). Graphical rewards were included in the Sustained Phonation Monitor (see sec 3.2.7). The inclusion of graphical rewards in all the modules, to make the aid appealing to young children, was a recurring suggestion in both the first and second assessment periods. We decided to place graphical rewards in the one module and assess their popularity.

At first it seemed the second assessment period of the CASTT did not yield as much as the first assessment period, but there was one important consequence of the second assessment period. This was the realization that the short-term evaluations could never provide information on the effectiveness of the CASTT as a speech aid. Another type of evaluation was necessary for that purpose.

The short-term evaluations (2 to 3 weeks per clinic) of the CASTT were very useful from an engineering perspective. Information was provided about the possibilities of new speech modules and the improvements needed to existing modules. However the short-term evaluations did not establish whether the CASTT contributed significantly to improvements in the clients' speech. Most clients only attend speech therapy on a weekly basis. The 2-3 week duration of the short-term evaluations barely gave the clients time to get used to using the aid. In addition the length of the evaluations was not long enough to observe any improvements in the clients' speech. It was decided that the CASTT should be placed in a clinic for a long period of time. This would give the therapists time to incorporate the CASTT into their training program and time for their clients to get used to using the CASTT. The results of these evaluations of the CASTT are outlined in the following section. However before this is done a few comments will be made about the limitations and strengths of the therapists' evaluations.

5.2 THE LONG-TERM EVALUATIONS OF THE CASTT

The next series of evaluations involved the CASTT being placed in a therapist's clinic for at least 2 months. These evaluations were called long-term evaluations. There were three therapists in this series of evaluations. The first long-term evaluation took place in late 1989, the next two in 1990. Two of the therapists had been involved in the second assessment period of the short-term evaluations. The clients of these therapists were children. The third therapist, who had not used the CASTT before, was in a private clinic, and her clients included both adults and children.

Apart from the addition of the Sustained Phonation Monitor module, the version of the CASTT that was placed in the clinics was the same as that in the second assessment

period of the short-term evaluations. An hour-long tutorial session was given to the therapist in private practice who was not familiar with the CASTT. It was the same tutorial as the one given to the therapists in the second assessment period of the short-term evaluations. No tutorial session was requested by the two therapists familiar with the aid.

Each therapist at the end of the long-term evaluation was required to write a written report on how useful they found the CASTT. They were asked to cite examples of clients using the speech modules. The intention of the evaluation was to get an indication of the CASTT's effectiveness as a speech aid. The CASTT was placed in a clinic for at least two months. In this time the therapists would have had a chance to observe if their clients' speech impediments were improving through the use of the CASTT.

As a consequence of the evaluations, we gained qualitative evidence suggesting that some of the CASTT's modules had helped improve the speech of some clients. The results of the evaluation will be discussed in the following section.

5.2.1 Three Therapists' Comments about the CASTT

The three therapists who were involved in the long-term evaluations will be referred to as TM, ET and TC respectively. TM and ET provided comments on all the modules of the CASTT. It is their comments that most of the following discussion will be based on. TC only provided comments on two modules. These will be mentioned in due course.

The Voice Pitch Tracker module displays a record of the variations of the pitch in an utterance vs. time. It is intended to be used for learning pitch control and to give an indication of pitch range, average pitch and pitch transitions in utterances. The module was the one most frequently used by TM, of it he said:

"children quickly learnt that their vocal behaviour affected what was happening on the screen. One child in particular achieved almost instant success in altering her pitch when this had previously been quite a difficult task for her."

TM said the Overlay option in the Voice Pitch Tracker, which enabled the client's contour to be directly overlayed on to the therapist's contour as the client spoke (see sec 3.2.1), was used in almost every session.

The Loudness Monitor module displays a record of the variations of loudness in an utterance vs. time. This module is intended to be used for loudness control and to provide an indication of loudness variations, loudness range, average loudness and loudness transitions in utterances. TM also found this module to be extremely helpful. He said:

"little training was required for the children to see exactly when they were speaking too softly or too loudly. The line which can be superimposed on the display helped the children see what was an appropriate volume level"

ET used the Voice Pitch Tracker, Loudness Monitor and Concurrent Loudness and Pitch modules often. She said that since the pitch and loudness of a speaker are linked it is necessary to work on both features together. ET would first work on the pitch and loudness features of a speaker separately, using either the Voice Pitch Tracker or Loudness Monitor modules. She then would work on both speech features together

using the Concurrent Loudness and Pitch module (this is the module which concurrently displays a record of the variations of pitch in an utterance versus time and a record of the variations of loudness in an utterance versus time).

ET cited an example in which she used both the Voice Pitch Tracker and the Loudness Monitor modules:

“the six year old boy I used [these modules] with had vocal nodules through voice misuse. [A] could not monitor himself for loudness or pitch and used a loud voice with the pitch being very low.”

She said it took one half hour session before A could relate the “patterns” on the screen to speech. It took three sessions for A to relate the auditory feedback (see sec 3.2.1) to the patterns on the screen. According to ET, A appeared to be benefiting from the aid. She said:

“learning was taking place as [A] certainly was speaking softly and controlled within the clinic setting. ”

The Sustained Phonation module was intended to be used to promote controlled phonation of a single sound. However neither ET nor TM used it in this manner. TM used the Sustained Phonation module for helping his clients learn breath control. He required his clients to sustain any sound for the duration of the plot (5.12 sec). If they were successful they were rewarded with one of the five cartoon sequences. TM did not cite an example of a client using the Sustained Phonation module. However he did say that he thought this module was an extremely effective module in helping children learn breath control. He also added that the children found the cartoon sequences motivating.

The Spectrogram module was intended to be used as an articulation corrector for any sound uttered in isolation, in a word or within a sentence. TM did not think the Spectrogram module was very good. He felt the module provided very little useful feedback to the client about their utterance. He also said that the Sustained Phonation module, used as it was originally intended, that is to promote the phonation of a single sound, provided little feedback to the client. He felt the plots for different sounds were not very distinctive. Interestingly whilst TM felt neither the Spectrogram or Sustained Phonation modules were effective articulation correctors, ET successfully used the modules to help remediate articulation errors.

ET used the Sustained Phonation module interchangeably with the Spectrogram module. She recognised that both modules were displaying the same information. ET used both modules as articulation correctors. She cited an example using the two modules to help a client, J, aged 8, set the sound [tʃ] at the word level and sustain it at the sentence level. J had a motor speech problem and pronounced many sounds incorrectly. She was pronouncing [ʃ] instead of [tʃ]. Using the display of the Spectrogram/Sustained Phonation module J was able to identify when she was substituting [ʃ] for [tʃ] in her speech, by the colour distribution in the plots. The pattern associated with the [tʃ] sound was red, white, blue and black; the pattern associated with the [ʃ] sound was white. Using this information J became able to correctly pronounce [tʃ] at the word and sentence level. ET, in agreement with TM, commented that the children found the cartoon sequences (in the Sustain Phonation module) motivating.

The Vocal Tract Shape module is an articulation corrector. It is intended to be used for correcting the articulation of vowels uttered in isolation. However neither ET nor TM felt the Vocal Tract Shape module was useful. TM felt the vocal tract shape reconstructions did not accurately represent, or even closely approximate the true shape of the vocal tract; ET reiterated this sentiment. Interestingly, the third therapist, TC, used the Vocal Tract Shape module with a client and observed improvements in the client's speech. The client, L (aged 35), was nasalizing her [a] sounds. L was placing her tongue very high in her mouth and reducing oral resonance. TC said that the Vocal Tract Shape module, used with other techniques, helped L establish the correct tongue placement.

The Fricative Monitor was the third articulation corrector in the CASTT. This module was intended to be used to indicate if frication had occurred in an utterance. Neither TM nor ET thought the Fricative Monitor module provided a consistent indication of frication. They both observed that [ʃ] was the only fricative the monitor consistently responded to. The third therapist, TC, used the module to help a client, M (aged 5), cease substituting [s] for [ʃ] and [ts] for [tʃ]. TC used the Fricative Monitor in conjunction with other techniques to get M to correctly pronounce [ʃ] and [tʃ].

5.3 THE CONSEQUENCES OF THE THERAPISTS' EVALUATIONS

The feedback from TM and ET about the Voice Pitch Tracker, Loudness Monitor and Concurrent Loudness and Pitch modules suggests that these modules were instrumental in the improvement in the speech of some of their clients. Both therapists noted improvements in their clients' speech as a result of using the modules. It is not so easy, however, to come to any clear cut conclusion about the Spectrogram, Sustained Phonation, Vocal Tract Shape and Fricative Monitor modules. For each of these four modules, one of the three therapists observed that the speech of some clients had improved after the use of the module. However, there was also at least one of the other therapists who questioned the worth of the module.

It is very tempting to dismiss those therapists who questioned the worth of the modules and claim they did so because they did not understand the displays. However the comments of TM and ET on the Vocal Tract Shape and Fricative Monitor modules were in fact to be expected. On looking more closely at these two modules it became apparent that neither module could be relied upon to provide useful information. An investigation of the Vocal Tract Shape module, carried out independently of the therapist evaluations, had come to the conclusion that the current reconstruction algorithm used in the module could not be relied upon to accurately reconstruct the vocal tract shape from a vowel sound. The investigation is outlined in sec 8.3.4.1. As a result, different methods of vocal tract reconstruction were reviewed in order to find a more accurate reconstruction algorithm. The review and the consequences are presented in chapter 8.

On closer investigation of the TMSboard, it became obvious that the Fricative Monitor module, in its current form, could never provide a consistent indication of frication. All the frequency components in a sound above 3 kHz were being attenuated due to the filter characteristics of the anti-aliasing filter on the TMSboard (see 4.2). Most of the spectra of fricatives have significant energy at frequencies higher than 3000

Hz. In fact only the spectra of the palatal fricatives ([ʃ,ʒ]) have a major concentration of energy at frequencies lower than 3000 Hz (see sec 7.4.1). Thus for virtually all the fricative sounds important information would be removed by the anti-aliasing filter on the TMSboard. The low-pass filtered fricative sounds would yield much fewer zero-crossings than expected since the zero-crossing rates are related to the spectral content (Rabiner and Schafer, 1978) and the high frequency components had been removed by the filter. Therefore the Fricative Monitor cannot consistently provide an indication of frication, since frication was said to occur if the zero-crossing rate was over a pre-set threshold. As a consequence of the Fricative Monitor being inadequate, a new fricative identification method was investigated. The results of this investigation are discussed in chapter 7.

Given that TM's and ET's assessment of the Vocal Tract Shape and Fricative Monitor modules was correct, the question arises as to why TC observed improvements in her clients' speech when she used the modules in therapy. TC used both modules in conjunction with other techniques. Therefore it could be argued that the clients' speech actually improved through the use of the other techniques, rather than the use of these modules. TC, herself, commented that the difference in the vocal tract shape between [a] and the nasalized [a] was minimal. This suggests that TC may have only been using the display to re-enforce to the client what she herself heard. There was no useful information about the utterance that could actually be obtained from the display.

Another reason for the improvements in the speech of TC's clients may have been that the clients were not benefiting from information obtained by the displays in the CASTT, but rather they were being motivated by the use of a novelty aid in their therapy. This is regardless of whether TC correctly interpreted the displays or not. This reasoning, however, means that it is therefore questionable whether the improvements observed in the clients' speech after use of the Loudness Monitor, Voice Pitch Tracker and Spectrogram modules can be attributed to the information obtained from the modules.

After some thought it became apparent that with only the evidence obtained from the long-term evaluation of the CASTT it is impossible to make any conclusive comments about the CASTT's effectiveness as a speech aid. This is not because the therapists were not of the same opinion about each module, nor is it because too few examples of client's speech improving through the use of the CASTT were cited. Even if many more examples of the client's speech improving through the use of the CASTT had been cited, these alone would not have been enough evidence for the CASTT being an effective aid.

Before a test like the long-term evaluation was performed the visual displays of the CASTT should have been assessed. It had not been established which significant phonetic and suprasegmental differences in speech could be detected from the displays of the CASTT. Nor had it been established how similar the displays of the CASTT are for utterances which do not have significant phonetic and suprasegmental differences. Once these were known we would know the speech errors for which the CASTT could be expected to have remedial potential.

At the time, there was no test available from which the displays of the CASTT could be assessed for their remedial potential. One has been developed by the author for that very purpose. The test is called the Visual Display Test. The following chapter,

chapter 6, outlines the development of this test. In addition it discusses the speech errors for which the CASTT modules can be expected to have remedial potential.

The Visual Display Test is only intended to establish exactly what information we can expect to obtain from the displays of the CASTT. The test will not be able to establish if the improvements observed in a client's speech are due only to the client being motivated by the use of a novelty aid. Only time and familiarity with the aid will establish that. Nor will the test establish if the therapist is only using the displays of the CASTT to re-enforce what they heard. However if a client can successfully use the CASTT alone using drill work, it suggests that there is information to be obtained from the screen. There is evidence that clients were able to use the CASTT alone for drill work.

In 1991 two therapists asked if they could have the CASTT in their clinics. For one therapist, ET, this was the third time she had had the CASTT. For the second therapist, KM, it was her first. No formalized method of assessment was employed during the time these therapists had the CASTT. At the end of a six months period the two therapists partook in a joint discussion with the author. The discussion centred mainly on the clients' interactions with the CASTT. Both ET and KM found that their clients were able to use the CASTT for speech drill. Once the client knew what to look for in the display, they could be left alone with the CASTT to practice. Both the therapists stressed that the CASTT was a great motivation for the clients. It gave them a degree of independence they did not have in traditional therapy. People with speech impediments rarely hear their impediment; they only hear what they intend to say. Those with speech impediments can resent being told their speech is incorrect. By using the CASTT the clients were able to "see" the speech was incorrect themselves. ET and KM found that the clients were far more amenable to their speech being identified as incorrect by the computer rather than themselves.

There is, therefore, some strong, albeit qualitative, evidence that the CASTT is an effective speech aid. The Loudness Monitor, Voice Pitch Tracker and Concurrent Loudness and Pitch modules appear to be very useful. The Spectrogram module was used successfully by ET (and incidentally by KM) to correct articulation errors. Finally both ET and KM were able to leave their clients alone with the CASTT to practise speech drill. However until we know what speech errors we can expect the CASTT to distinguish, we can only conjecture that the CASTT is an effective aid.

At the beginning of this chapter it was stated that in order for the CASTT to be an effective aid it must be developed with continual feedback from the speech therapists. This chapter has shown how the therapists' comments have resulted in improvements to the CASTT's modules and to the addition of 2 new modules. Through observation of the therapists using the CASTT I was able to make the CASTT easy to operate and user-friendly. The fact that the therapists only required a one hour long tutorial before using the system gives testimony to this; especially considering most of the therapists had not used a computer before. Observing the therapists use the CASTT also enabled me to observe difficulties that people not familiar with the CASTT had; these difficulties would have gone unnoticed if the CASTT had not been used in a clinical situation. If I had not seen how common it was for there to be a large time delay between the time the Loudness Monitor and Voice Pitch tracker modules were placed in the active mode and the time someone spoke into the microphone, the modules would not have been altered so the initial silence was not plotted (see sec 5.1.1.1).

By the second assessment period it became apparent that some of the speech therapists were not assessing the CASTT critically. The versions of the CASTT placed in the clinics, as mentioned earlier, were not the optimum version. It was expected that there would be some bugs in the modules. The therapists were all told that the displays of the CASTT modules might not necessarily impart correct information, indeed the need to check for bugs was one of the reasons for the evaluations. However it was very difficult to get some of the therapists to offer any criticism of the CASTT; this was partly, I think, out of politeness. If the CASTT is to be evaluated effectively, the therapists themselves must evaluate it critically. However the therapists who evaluated the CASTT were skilled practitioners and not, for the most part, skilled researchers. Whether they evaluated the CASTT critically or merely accepted the CASTT for what it was, was essentially the luck of the draw. Having said that, the vast majority of the therapists who did evaluate the CASTT did provide some constructive criticism.

If the CASTT had been developed by a research team comprising both engineers and speech therapists, similar to the BBN, VSA, ISTR, IBM-SpeechViewer and John Hopkins aids (see sec 2.3), then the above mentioned problem would not have occurred. There was one advantage, though, of there not being a pre-set group of therapists associated with the CASTT project. There was no danger the CASTT was only tailor-made for a particular group of therapists. Many different therapists have had input into the CASTT project. Since most of the comments by the therapists about the CASTT have been very positive, it bodes well for the potential of the CASTT as an effective speech aid.

CHAPTER 6

AN EVALUATION OF THE CASTT VISUAL DISPLAYS

6.1 INTRODUCTION

The evaluations of the CASTT described in the previous chapter have been of a qualitative nature. They led to a number of significant improvements in the software packages, the “user friendliness” of the help manuals and software menus, and the manner in which the speech plots were presented in the CASTT speech analysis modules. The feedback from the therapists strongly suggested that the CASTT was instrumental in the improvement of some of their clients’ speech, (see sec 5.2.1). However before any definitive comment can be made on the worth of the CASTT as a speech aid, it is necessary to establish exactly what information one can expect to obtain from the displays. Therefore the final consequence of the speech therapists’ evaluations of the CASTT was the realization of the necessity to assess the displays of the CASTT (see sec 5.2.1). It has to be established which significant phonetic and suprasegmental differences in speech can be discriminated from the displays of the CASTT and it has to be established how similar the displays of utterances with no significant phonetic or suprasegmental differences are. No test currently exists to do this.

It is interesting that the past developers of Visual Speech aids reviewed in sec 2.2 and 2.3 never developed such a test. In the past it had been usual practice, when a visual speech aid was at the stage that the CASTT is now, to perform an evaluation of the aid, obtaining data on the aid’s effectiveness in improving clients’ speech. The aids were generally assessed in one of two ways. The first was done by comparing the recorded speech of clients before and after a period of therapy using the aid (Pronovost *et al.*, 1968; Stark, 1971; Boothroyd *et al.*, 1975; Bristow *et al.*, 1981; Kewley Port *et al.*, 1987a). The other method of assessing aids was comparing the improvement in speech of two groups of clients, all with the same type of speech disability. One group would use traditional methods to ameliorate the disability; the other group would use the computer-based aid (Stark, 1972; Povel, 1974b; Osberger *et al.*, 1981; Bate *et al.*, 1982; Pardo, 1982; Arends *et al.*, 1991).

The assessments of the visual speech aids invariably established that clients who use a visual speech aid in speech therapy corrected their speech impediments more quickly than those clients who used conventional methods. However reasons were never given as to why this should be the case. The noted improvements may have been due to the therapists and the clients correctly interpreting the plots or merely due to the clients being motivated by the special treatment they were getting in using a novelty aid. In addition it was never established from these assessments, whether therapists use the visual information on the screen to assess the client’s speech, whether they could use the display to show the client the difference between what was spoken and what was

wanted, or whether they use the information merely to reinforce what they have already heard.

To date there has been very little research on whether the therapists and clients actually understand the plots of visual computer-based speech therapy aids, or how to test that the information obtained from the plots will be useful in speech therapy.

The purpose of this chapter is to present my visual display test. The test is called the VDT. It has been designed to be used to assess the potential of a computer-based speech therapy aid like the CASTT. The test was performed on the CASTT; the results will be presented and discussed. The test evolved from research done by Braeges and Houde (1982) and a preliminary test that I developed. Before presenting the VDT, the research which led to its development will be outlined and discussed.

6.2 ASSESSMENT OF VISUAL DISPLAYS OF COMPUTER-BASED SPEECH THERAPY AIDS

A few researchers, such as Bargstadt *et al* (1978) and Maki *et al* (1981), have made some attempt at assessing the displays of specific visual speech aids. Bargstadt *et al* (1978) investigated the success rate of five trained subjects in identifying the eight English fricatives from visual patterns of the Video Articulator (the patterns were similar to Lissajous figures). Maki *et al* (1981) were interested in the plots of the Speech Spectrographic Display. They investigated whether hearing-impaired subjects could identify which of two utterances was the better pronounced from the plots and whether the subjects could correctly relate the visual patterns to selected articulatory events (Maki *et al.*, 1981).

Whilst Bargstadt *et al*'s assessment checked the consistency of the visual patterns for specific sounds it did not establish whether the aid was able to detect errors in speech. The assessment of Maki *et al* did establish this and what is more it sought to establish that the subjects could relate the visual patterns to speech characteristics. An evaluation of this sort could be carried out on the displays of the CASTT. One substantial drawback would be the number of speech errors for which the CASTT would have to be tested.

The errors in speech fall into three categories, articulation, phonation and fluency (Skinner and Shelton, 1978). Research cited by Braeges and Houde (1982) has demonstrated that there are around 600 types of common speech errors made by the hearing-impaired. No known set of common speech errors for people with speech disorders who are not hearing-impaired has been compiled. It will be assumed that this set can be represented by the common speech errors made by the hearing-impaired. The CASTT should be able to distinguish all of the 600 common speech errors because it has modules that plot suprasegmental and paralinguistic features of speech; the Voice Pitch Tracker, the Loudness Monitor Module, the Concurrent Pitch and Loudness Module and the Sustained Phonation Module, and modules that plot the phonetic features of speech; the Spectrogram Module, Fricative Monitor, Vocal Tract Shape and Lissajous Figure module (this is the proposed new module of the CASTT which will display Lissajous figures, see sec 4.8.2). To test the displays of the CASTT for the common 600 speech errors would take a considerable amount of time.

6.2.1 The Short Test Of Elementary Error Discrimination

There is one visual display test proposed in the literature that could be carried out on the CASTT. Braeges and Houde (1982) basically extended Maki *et al's* ideas and have developed a test called the Short Test of Elementary Error Discrimination (STEED). The STEED utilises a list of 29 elementary speech errors, which are representative of the 600 common speech errors. Each of these 29 errors is highlighted by a target utterance and an error utterance. The pairs of target and error utterances differ in only one aspect of speech (e.g. articulation intensity, speech timbre, etc). The 29 elementary errors and the speech pairs (made up of a target and error utterance) that exemplify the errors are listed in tables 6.1 and 6.2.

The Elementary Error list can be subdivided into six sets. Each set contains a certain type of speech error. The sets are : the Articulatory Intensity Set which represents errors resulting in intensity change due to misarticulation (as opposed to suprasegmental aspects); the Voiced/Unvoiced Set which represents errors in the voiced/unvoiced distinction; the Nasality Set which represents errors in nasality; the Articulation Substitution Set which represents errors in phoneme substitution; the Suprasegmental Set represents errors in suprasegmental aspects of speech; and finally the Speech Timbre Set which represents errors in speech timbre.

Braeges and Houde believed that, for each elementary error, if a user could differentiate between the target and error utterance pairs which exemplified it, using a visual speech aid display, then the aid could be used effectively in therapy for all the common speech errors (Braeges and Houde, 1982).

However, Braeges and Houde's STEED neglected to test for the consistency and repeatability of an aid's visual pattern when two utterances are judged to be the same. Certainly it is imperative that the difference between two utterances which differ in one aspect of speech, such as a phoneme or pitch, (henceforth referred to as a "**different-speech pair**") should be immediately noticeable from the plots of a visual speech aid. It is also desirable that two utterances with the same phonetic transcription and suprasegmental features (henceforth referred to as a "**same-speech pair**") be seen as more similar than any different-speech pair. Of course, this criterion of "sameness" cannot be satisfied by all same-speech pairs, because there will always be some different-speech pairs straddling the boundary between acceptable and unacceptable utterances which are more similar than the same-speech pairs spanning the region of acceptable utterances.

6.3 THE PRELIMINARY TEST OF THE CASTT'S VISUAL DISPLAYS

The first test developed to assess the displays of the CASTT was adapted from the STEED. In the test, the plots resulting from both same- and different-speech pairs were shown to participants. From the plots the participants were required to identify the speech pair type (i.e. a same- or different-speech pair). The same- and different-speech pairs were obtained from the elementary error lists.

| Elementary Error | Description Of Error | Target/Error Combination |
|-------------------------------|---|--|
| Articulation Intensity Set | | |
| EE-1 | Detects the release of a complete articulatory closure | boo/boot |
| EE-2 | Detects the release of a partial articulatory closure , and the change in the envelope level due to a change in voicing | allay/away |
| EE-3 | Distinguishes an initial unvoiced plosive from a voiced plosive | two/do |
| EE-4 | Distinguishes sounds which differ in duration and in average intensity | poppa/pop |
| EE-5 | Distinguishes sounds whose speech envelopes differ in initial or final rate of change | ban/ran |
| Voiced/Unvoiced Set | | |
| EE-6 | Distinguishes sustained unvoiced sounds from voiced sounds | I/shy |
| EE-7 | Detects the occurrence of very short duration unvoiced sounds | oh/toe |
| EE-8 | Distinguishes sustained voiced fricative sounds from voiced sounds | Ann/van |
| EE-9 | Distinguishes sustained voiced fricative sounds from unvoiced fricative sounds | Sue/zoo |
| Nasal Quality Set | | |
| EE-10 | Distinguishes errors in velar state (open versus fully closed) | me/bee |
| EE-11 | Detects errors in the timing of the velar closure or release in nasal sounds. | me/mbee |
| Articulation Substitution Set | | |
| EE-12 | Distinguishes s/ʃ substitutions | see/she |
| EE-13 | Distinguishes far-neighbour front vowel substitutions | be/ba |
| EE-14 | Distinguishes far-neighbour back vowel substitutions | do/door |
| EE-15 | Distinguishes near-neighbour front vowel substitutions | Ed/add |
| EE-16 | Distinguishes near-neighbour back vowel substitutions | fought/foot |
| EE-17 | Distinguishes substitutions of diphthongs | down/don |
| EE-18 | Distinguishes addition of glides | use (verb)/ ooze |
| EE-19 | Distinguishes errors differentiated by very small spectral change | way/ray |
| Suprasegmental Set | | |
| EE-20 | Distinguishes the difference in syllable pitch corresponding to a difference in stress | spot ¹ /spot ² (1 and 2 differ in pitch by a musical whole tone) |
| EE-21 | Distinguishes differences in terminal pitch contour | now ?/now ! (emphasized) |
| EE-22 | Distinguishes syllable loudness and syllable duration | contract(noun)/contract(verb) (pitch monotone) |

Table 6.1. The elementary errors numbers 1 to 22 and the target/error combinations which are an example of that error, from Braeges and Houde (1982) .

| Elementary Error | Description Of Error | Target/Error Combination |
|--|---|--|
| Speech Timbre Set | | |
| For EE-23 to EE-29 use any short phrase, vary trials by test measure only | | |
| EE-23 | Distinguishes error in average pitch | Normal pitch range/Half an octave higher |
| EE-24 | Distinguishes error of excessive loudness | Normal loudness/ Too loud |
| EE-25 | Distinguishes error of insufficient loudness | Normal loudness/ Too quiet |
| EE-26 | Distinguishes error of too slow speaking rate | Normal rate / Too slow |
| EE-27 | Distinguishes hypernasality errors | Normal Nasality / Hypernasality |
| EE-28 | Distinguishes error due to a breathy voice | Normal voice / Breathy voice |
| EE-29 | Distinguishes error due to a tense voice | Normal voice / Tense voice |

Table 6.2. The elementary errors numbers 23 to 29 and the target/error combinations which are an example of that error, from Braeges and Houde (1982) .

6.3.1 The Modification Of The Speech List Which Represents The Elementary Errors

A few modifications had to be made to the target and error utterance pairs in the Elementary Error list, for it to be used with the CASTT. The original target and error utterance pair examples had been designed for American English. For the examples to be true for New Zealand English one change was necessary, this being for elementary error EE-14. The error was "far-neighbour back vowel substitution", and the target/error speech pair example was given as "do/da". In New Zealand English "da" is pronounced with a central vowel rather than a back one. Thus the speech pair example for elementary error EE-14 was changed to "do/door". The modified and unmodified speech pairs from the elementary error lists in table 6.1 and table 6.2 will henceforth be referred to as speech list NZ-SL1. The phrase selected for illustrating elementary errors EE-23 to EE-29 was the phonetically balanced sentence "Grape juice and water mix well".

The Elementary Error list, and hence NZ-SL1, was designed to be used to assess the visual aids which have time-plots, that is displays which show a record of speech characteristics as they varied with time. Hence plots of the Voice Pitch Tracker module, the Loudness Monitor module, the Concurrent Pitch and Loudness module, the Spectrogram module and the Sustained Phonation module were assessed using same- and different-speech pairs from NZ-SL1.

The other three modules - the Fricative Monitor, the Vocal Tract Shape and the Lissajous Figure modules - display speech characteristics for an instant of speech only, that is they are current-value-plots. Thus a second speech list had to be compiled to test these modules; this will henceforth be called NZ-SL2. This list contained only twelve elementary errors, which are listed in table 6.3. All the speech errors that involved aspects of time were removed. Thus elementary errors EE-1, EE-2, EE-11 and EE-17 were removed because they involved errors in co-articulation and elementary errors EE-4, EE-5 and EE-7 were removed because they involved errors in sound duration. The elementary errors of the suprasegmental aspects of speech and speech timbre (EE-20 to EE-29) were removed because the three modules were only intended to detect phonetic

| Elementary Error | Description Of Error | Target/Error Combination |
|-------------------------------|--|--------------------------|
| Articulation Intensity Set | | |
| EE-3 | Distinguishes an initial unvoiced plosive from a voiced plosive | [t]/[d] |
| Voiced/Unvoiced Set | | |
| EE-6 | Distinguishes sustained unvoiced sounds from voiced sounds | [i]/[j] |
| EE-8 | Distinguishes sustained voiced fricative sounds from voiced sounds | [æ]/[v] |
| EE-9 | Distinguishes sustained voiced fricative sounds from unvoiced fricative sounds | [s]/[z] |
| Nasal Quality Set | | |
| EE-10 | Distinguishes errors in velar state (open versus fully closed) | [m]/[b] |
| Articulation Substitution Set | | |
| EE-12 | Distinguishes s/ʃ substitutions | [s]/[ʃ] |
| EE-13 | Distinguishes far-neighbour front vowel substitutions | [i]/[a] |
| EE-14 | Distinguishes far-neighbour back vowel substitutions | [u]/[ɔ] |
| EE-15 | Distinguishes near-neighbour front vowel substitutions | [e]/[æ] |
| EE-16 | Distinguishes near-neighbour back vowel substitutions | [ɔ]/[ʊ] |
| EE-18 | Distinguishes addition of glides | [j]/[u] |
| EE-19 | Distinguishes errors differentiated by very small spectral change | [w]/[r] |

Table 6.3. The list NZ-SL2 with the phonemes given as target/error combinations. The brief explanation of the errors is from Braeges and Houde (1982) .

features of speech.

Since the second group of modules only displayed brief time-invariant speech characteristics, the speech pair examples which represented the errors had to be single phones. These were derived from the words and sounds in NZ-SL1, taking into consideration the elementary error highlighted by the original speech pair. For example, from table 6.1 “Sue/zoo” exemplified the speech error “ substituting unvoiced fricatives for sustained voiced fricatives”, so the target/error utterance pair in NZ-SL2 became “[s]/[z]”. The reduction process was straight forward, with the exception of the target/error combination to exemplify EE-6. The description for this error was “ distinguishes sustained unvoiced sounds from voice sounds”. The sustained voiced sound used in NZ-SL1 was the diphthong [ai]. Since diphthongs involve a time component they cannot be observed in the plots of the Fricative Monitor, the Vocal Tract Shape or the Lissajous Figure modules. To overcome this problem, the sustained voiced monophthong [i] was used as the target utterance instead of [ai] Thus the target/error utterance pair in NZ-SL2 for elementary error EE-6 was [i]/[j]. NZ-SL2 is a subset of NZ-SL1.

6.3.2 Obtaining The Pre-Recorded Speech For The Preliminary VDT

For each target/error combination (denoted by X/Y say) from the elementary error lists four utterances were pre-recorded, two utterances of X and two of Y . From these utterances two same-speech pairs (X_1X_2 , Y_1Y_2) and two different-speech pairs

(X_1Y_1, X_2Y_2) were made (a further two different-speech pairs are possible but were not used). From each of these utterances plots were obtained for each visual display type.

Pre-recorded speech was used to ensure that the variations in the visual patterns were due to the utterances themselves rather than because the speech was obtained from many different speakers or one speaker on many different occasions. Also it ensured the varying background conditions were limited, and it enabled all the speech used in the test to be checked for correct pronunciation. Finally, because the VDT was a visual test, pre-recording ensured that the participants could not overhear the utterances as they were being spoken.

The words and sounds of the two speech lists were recorded by a male speaker. The speaker was not professionally trained but he spoke clearly, had no speech impediments and spoke with a New Zealand accent.

A randomly ordered list of words and sounds was made from the target/error utterance pair examples of elementary error EE-1 through to EE-19 from NZ-SL1, each utterance appearing on the list twice. Another randomly ordered list was made from the target/error phones in NZ-SL2. The words/sounds/phones were read out in groups of three and recorded. Each word or sound was spoken with equal emphasis. To get the correct contrast between the speech pairs exemplifying elementary errors EE-20 to EE-29, the words or phrases in the target/error combinations were recorded consecutively. Once again two versions of each word and phrase were recorded.

The speech was recorded and then stored as digitized samples. To do this it was amplified and passed through a 4.5kHz 8 pole elliptic filter into a 16 bit A/D. The filter and A/D were part of an Antex Electronic SX-8 Digital Audio Processor (the SX-8 board resided in an IBM-PC AT). The digitized speech was stored on the hard disk of the IBM-PC after the silences had been edited out. The speech was sampled at 10 kHz. All recording was done in a quiet room. Each of the digitized words, sounds, sentences and phones were listened to by myself to ensure the stored speech was correctly pronounced, correctly labelled, of good quality, not truncated, and all digitized at the right sampling rate. Several utterances had to be recorded again.

6.3.3 Preparing The Visual Displays For The Preliminary VDT

Though there are eight speech analysis modules in the CASTT, there are only six distinct visual display types. These are pitch contours, loudness contours, spectral content, frication content, vocal tract shape estimation and Lissajous figures.

All but one of the current modules of the CASTT have two plots. The exception is the Fricative Monitor. In the normal operation of the CASTT one plot is for the results of the client's speech and the other for a reference plot provided by the therapist. The plots for each visual display type in the preliminary VDT were kept in the same format as for the equivalent speech analysis module of the CASTT, e.g. keeping the scale and positioning of the plots the same. In the case of the CASTT's Fricative Monitor, which has only one plot, a second plot had to be added for the preliminary VDT. The actual display format of the proposed Lissajous Figure module has not yet been decided upon. In the Preliminary VDT the size of the Lissajous figure plots were made to be the same size as the vocal tract reconstruction plots.

All the plots were calculated using the same signal processor as that used by the CASTT, the TMS32010. All the plots in the VDT were pre-calculated from pre-

recorded speech. Due to this, minor changes were needed in the data handling sections of the TMS32010 software for the six display types. The sampled speech was placed into the TMS32010 for processing from a file rather than straight from the A/D as in the real-time aid (see sec 4.2). Since the speech processed by the TMS32010 for the VDT was passed through the low-pass filter SX8 board not on the TMSboard it had a bandwidth of 0-4.5kHz not 0-3.0kHz (see sec 4.2). In addition the pre-recorded speech was stored as a 16 bit number. The TMS32010 software was designed to process 12 bit speech only (the range of the A/D on the speech processing board of the CASTT, see sec 4.2). Before the pre-recorded speech was passed into the TMS32010 it was scaled to 12 bits.

6.3.4 The Participants Of The Preliminary VDT

In this preliminary investigation there were nine participants, four women and five men. All were staff or postgraduate students of the Department of Electrical and Electronic Engineering at the University of Canterbury except one, who was a school leaver working in the aforementioned department over the summer vacation.

6.3.5 The Results Of The Preliminary VDT

The intention of the preliminary VDT was to establish that, for each elementary error exemplified by a target and error utterance pair from either NZ-SL1 or NZ-SL2, there is at least one visual display type in the CASTT from which same- and different-speech pairs can be distinguished. Since the display format allowed only one pair to be displayed at a time, the test relied on the participants to provide their own threshold of difference between "same" and "different". The participants were *not* told whether their responses were correct or incorrect. However they *were* told which target/error combination the pair was obtained from (for example they were told the plots were from any same- different-speech pair combination of "boo/boot"). They had to say whether the plots showed a same-speech pair or a different-speech pair.

The results indicate that even under these conditions the participants did considerably better than they would have by pure guessing. In table 6.4, the column labelled "8up level" gives the numbers of correct responses by eight or more of the nine participants. Similarly, the "7up level" gives the numbers of correct responses by seven or more participants. The first row of the table gives the number of correct responses for identifying all four speech pairs resulting from a target/error combination. The second and third rows give the number of correct responses for identifying both different-speech pairs or both same-speech pairs respectively. In brackets is the number of correct responses for identifying the specified set of speech pairs expected if all the same- and/or different-speech pair decisions were made randomly, rounded to the closest integer. Thus, the top left entry means that for three out of the twenty nine speech errors, each of the four speech pairs was identified correctly by eight or more of the participants and if all the same- and different- speech pair decisions were made randomly then we would expect that for none of twenty nine speech errors would the four speech pairs be identified correctly by eight or more of the participants.

The probabilities of random guessing were calculated assuming the binomial distribution; these values are listed in table B.1 in appendix B.1.

| Successful Identification | NZ-SL1 | | NZ-SL2 | |
|-----------------------------|------------------|------------------|-----------------|-----------------|
| | 8up level | 7up level | 8up level | 7up level |
| all four speech pairs | 3 out of 29 (0) | 9 out of 29 (0) | 3 out of 12 (0) | 4 out of 12 (0) |
| both different-speech pairs | 18 out of 29 (0) | 25 out of 29 (7) | 9 out of 12 (0) | 9 out of 12 (1) |
| both same-speech pairs | 5 out of 29 (0) | 13 out of 29 (7) | 6 out of 12 (0) | 6 out of 12 (1) |

Table 6.4. The results of the preliminary VDT. The amount in brackets is the successful identification we expect (rounded to the closest integer) if all the same/different decisions were made randomly for the CASTT.

It is not surprising that the participants in this preliminary test were biased in their responses, since they themselves had to provide the threshold between “same” and “different”. For the plots resulting from the NZ-SL1 list, 62% of the responses were “different-speech pair” while for the plots resulting from the NZ-SL2 list the percentage was 56.3%, see table B.2 and table B.3 in appendix B.1. The bias also shows in table 6.4 where there is considerably greater success in correctly identifying the different-speech pairs. If the responses of correctly identifying different-speech pairs are analysed then we find that 57 of the possible 82 ($=2 \times (29+12)$) were unanimously (9 out of 9) selected correctly with at least one visual display type, and only one of the remaining 24 was identified correctly by less than half of the participants on all of the modules.

6.3.6 Discussion Of The Preliminary VDT Results

The preliminary VDT was carried out with a minimal disturbance to the CASTT so the plots would be seen by the participants in the same format as those seen by therapists and clients. Only in the case of the Fricative Monitor module did an additional plot have to be provided because just one plot is normally used with this module. Taken at face value, the results shown in table 6.4 could be taken to portray the CASTT as a poor aid which could not be used to distinguish more than a few elementary speech errors. This is at odds with the reports of the speech therapists who have used the aid (see sec 5.2.1). The reason for this apparent anomaly becomes clear as soon as we consider the nature of the preliminary VDT.

When a participant in the test judges a speech pair to be “different”, there is no doubt that the CASTT plots are showing a difference between the two utterances. For a “same” response we do not know whether the response is due to no difference being observed or because the participant has deemed the difference small enough to count as “same”. From this it can be seen that the “different” responses to a different-speech pair are testing the CASTT modules for their ability to distinguish speech errors but the “same” responses are testing the participants’ ability to set a threshold on difference for sameness. The preliminary VDT conflates the CASTT errors with participant errors.

The success of the CASTT in displaying differences between the plots for utterances in different-speech pairs is indicated by the numbers given above for the majority judgements of “different” for different speech pairs. Only one of the two different-speech pairs from the elementary errors failed to be judged as different by the majority of the participants using at least one of the modules.

With this support for the CASTT, we must put the blame for the relatively poor results given in table 6.4 on the preliminary VDT itself. In the next section we will

present the revised VDT. In this test only the plots of the CASTT will be tested, the participants' familiarity with the CASTT and ability to interpret the plots will not be a consideration.

6.4 THE VISUAL DISPLAY TEST (VDT)

The Visual Display Test (VDT) was a test which tests the visible speech aid, the CASTT, both for its ability to separate different-speech pairs and for its ability to display same-speech pairs as less different than different-speech pairs. In the VDT the participants were presented with the plots of three utterances. Two were from a same-speech pair and the third an error utterance. These sets of three plots will henceforth be called a plot-set. The participants selected which two of the three plots in the plot-set were the most similar. In doing this they only needed to judge the relative differences between the pairs of displays. They did not need to use any thresholds of similarity as was required by the preliminary VDT.

The utterances used in the VDT were drawn from the target/error combinations in the speech lists NZ-SL1 (see table 6.1 and 6.2) and NZ-SL2 (see table 6.3) Four plot-sets arise from each of the target/error combinations in the speech lists. Using the notation in sec 6.3.2, for a target/error combination X/Y , the plot-sets viewed were $X_1X_2Y_1$, $X_1X_2Y_2$, $X_1Y_1Y_2$ and $X_2Y_1Y_2$. The order, in which the plots of the utterances appeared within the plot-set, was random. Ideally the displays that look the most similar will be those from the same-speech pairs, X_1X_2 and Y_1Y_2 . For each elementary error we sought to establish that there was at least one display type in the CASTT in which the same-speech pairs in each of the four plot sets could be identified.

6.4.1 The Preparation Of The VDT

6.4.1.1 Obtaining The Pre-Recorded Speech For The VDT

Pre-recorded speech was used in the VDT, for the same reasons as it was in the preliminary VDT. The same male speech was used. In addition the words and sounds of NZ-SL1 and NZ-SL2 were recorded by a female speaker. The speaker was not professionally trained but she was actively involved in drama. She spoke clearly, had no impediments and spoke with a New Zealand accent.

The recording of the female speaker was done in exactly the same manner as it was for the male speaker, using the same recording system (see sec 6.3.2 for details). As with the male speech each digitized word, sound, sentence and phone was checked by the author to ensure the stored speech was correctly labelled, correctly pronounced, of good quality, not truncated and all digitized at the right sampling rate.

Several of the sounds had to be re-recorded. Unfortunately the speaker had left for overseas. The sounds were re-recorded using my voice. To ensure relative consistency both the same-speech pairs were re-recorded, even if only one of the utterances required re-recording. The sounds that were re-recorded were foot, fought, see, she, boo, boot and the sentence "Grape juice and water mix well" spoken at high pitch, from NZ-SL1 and the phones [i,æ,v,s,b] from NZ-SL2.

6.4.1.2 Preparing The Visual Displays For The VDT

The visual display types which were tested in the VDT were the same as those tested in the preliminary test, i.e. the pitch contours, the loudness contours, spectral content, frication content, vocal tract shape estimation and Lissajous figures. The plots were calculated in the same manner as for the preliminary VDT (see sec 6.3.3). However unlike the preliminary VDT, this VDT displayed three plots on the screen simultaneously.

6.4.1.3 The Elementary Errors For Which The Display Types Were Tested

In the VDT the pitch contours, loudness contours and spectral content display types were not tested for every elementary error exemplified by NZ-SL1. Similarly the fricative content response, vocal tract shape estimation, and Lissajous figures display types were not tested for every elementary error exemplified in NZ-SL2. This was because each of the display types was not intended to detect every speech error (the types of speech errors each module was intended to detect was discussed in chapter 3). For example the Vocal Tract Shape module was designed only to reconstruct the vocal tract shape of vowel sounds (see sec 3.2.6). For this reason there was no point in testing for elementary errors EE-3, EE-9, EE-12.

Table 6.5 lists the elementary errors for which each display type was tested. The elementary errors which were tested with both the time-plot and current-value-plot display types are indicated with an asterisk (e.g. EE*-2). The rest of the elementary errors involve some component of time and are only tested with the time-plot display types. The key for the different display types used in table 6.5 is listed beneath it.

6.4.2 The Presentation Of The VDT

The VDT was presented to the participants in a computer package which ran on an IBM-PC AT. For each of the display types the participant viewed a series of plot-sets. The plots within the plot-sets were either presented side by side (the spectral content, Lissajous figures and vocal tract shape estimation display types) or one beneath the other (the pitch contours, loudness contours and frication content display types). The participants were required to enter into the computer via the keyboard which two plots, in each plot-set, they felt were the most similar. The labels of the two plots selected were then highlighted on the screen. If these were the plots the participants intended to select then the participant accepted them and a new plot-set appeared and the decision process repeated. However, if they were not, the participant was able to make the selection again.

The order in which the plot-sets were presented, for each display type, was selected at random but it was the same for all the participants. The VDT was divided up into two sessions of equal length. In each session each of the six display types was tested. In the first session the order in which the display types was tested was spectral content using female speech, frication content using female speech, vocal tract shape estimation using male speech, loudness contours using male speech, pitch contours using female speech and Lissajous figures using female speech. In the second session the order was spectral content using male speech, loudness contours using female speech, frication content using male speech, pitch contours using male speech, Lissajous figures using

| Speech Element | Visual Display Type |
|-------------------------------|---------------------|
| Articulatory Intensity Set | |
| EE-1 | L S |
| EE*-2 | P L S F V J |
| EE-3 | P L S |
| EE-4 | P L S |
| EE-5 | L S |
| Voiced/Unvoiced Set | |
| EE*-6 | P L S F V J |
| EE-7 | S |
| EE*-8 | P L S F V J |
| EE*-9 | L F J |
| Nasality Set | |
| EE-10 | L S F V J |
| EE-11 | P L S |
| Articulation Substitution Set | |
| EE*-12 | S F J |
| EE*-13 | L S F |
| EE*-14 | P L S V J |
| EE*-15 | P L S F V |
| EE*-16 | S |
| EE-17 | P S |
| EE*-18 | S F V J |
| EE*-19 | P S F V J |
| Suprasegmental Set | |
| EE-20 | P L |
| EE-21 | P L S |
| EE-22 | P L S |
| Speech Quality Set | |
| EE-23 | P L |
| EE-24 | L |
| EE-25 | L |
| EE-26 | P L |
| EE-27 | P L S |
| EE-28 | P L S |
| EE-29 | P L S |

Table 6.5. The elementary errors for which each of the six display types were tested. P stands for the pitch contours, L for loudness contours, S for spectral content, F for frication content, V for vocal tract shape estimation and J for Lissajous figures. The elementary errors which were tested with both the time-plot and current-value-plot display types are indicated with an asterisk (e.g. EE*-2).

male speech and vocal tract estimation using female speech.

There was no particular reason for the order of presentation for the display types apart from the spectral content displays. The spectral content response display types were assessed first in both sessions because these plots took the longest to appear on the screen. It was thought it would minimize the participants' frustration in doing the VDT if the most time consuming display type was assessed first.

The entire VDT took between one and a half and two and a half hours to complete, the time varying from participant to participant. Most people did the VDT in two separate sessions but several completed the VDT in one sitting. The participants viewed 720 plot-sets in total.

6.4.3 The Participants Of The VDT

There were 31 participants who did the VDT. The participants were recruited in three ways: some were friends of mine, some were alerted to the experiment via a notice on the departmental computer network and some were informed about the experiment in an undergraduate electronics laboratory. Seven of the participants were women and twenty four were men. Sixteen of the participants were either staff or postgraduate students of the Department of Electrical and Electronic Engineering, seven were undergraduate students in the same Department, and eight were trained in other disciplines. All the participants were university educated. The ages of the participants ranged from eighteen through to the mid-thirties, with the exception of one participant who was in his mid-fifties.

There was a payment of \$15 available for those who did the test but only twenty people accepted payment.

6.4.4 Remedial Potential

Before any results of the VDT are presented and discussed, it is necessary to introduce the concept of "remedial potential" and to define it. For the CASTT to have remedial potential for a particular elementary error, the therapist and client must be able to distinguish target and error utterances which exemplify the error. For this to be true, the plots of the target utterances must look different from the plots of the error utterances. In addition the plots from a series of utterances with the same phonetic transcription and suprasegmental features must be seen as more similar than the plots of utterances with different phonetic transcriptions or suprasegmental features.

However there will be some situations in which the plots of a target and an error utterance straddling a boundary are seen as more similar than the plots of a pair of target utterances which lie in the region of the particular phonetic transcription. In addition, due to the subjective nature of perception we cannot assume that the boundary will remain in the same place for each participant. Hence we can never assume that the plot of a particular target utterance and the plot of a particular error utterance will always be perceived as different. Therefore less than 100% success in distinguishing target and error utterances from plots must be accepted.

We can see from curve (a) in figure 6.1 that there is a very strong agreement between participants' perception of what is visually similar. Figure 6.1 (a) is the cumulative graph of the number of times in which z or more participants selected the same two

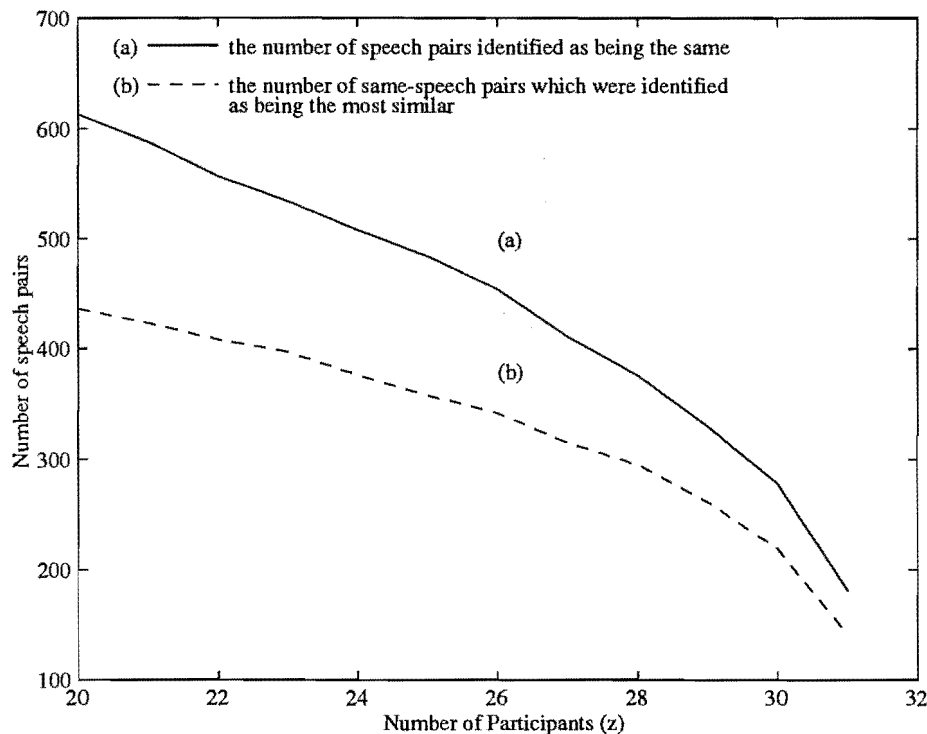


Figure 6.1. The plots for (a) the number of times in which z or more participants selected the same two displays as being most similar (regardless of whether the displays were obtained from a same- or -different speech pair), and (b) the number of same speech pairs correctly identified by at least z participants, where $z = 20$ to 31.

plots as being most similar, where $z = 20$ to 31, regardless of whether the plots were of a same- or different- speech pair.

It can be seen from curve (a) in figure 6.1 that in 181 of the 720 plot-sets (25.1 %) all of the participants selected the same two plots as being the most similar. Incredibly 90.3 % or more of the participants (i.e. 28 or more) selected the same two plots as being the most similar in 376 plot-sets (52.2 %). This is a very encouraging result. Most of the participants had no prior interaction with the aid. Despite this they appeared to use comparable visual cues in order to make their decisions as to which two displays were the most similar. The visual cues the participants used would vary for each display type. Length, shape, height and colour would have been some of the cues used.

For the displays from visual aids like the CASTT to be useful there must be an agreement in what people find to be similar and what they find to be different. From curve (a) in figure 6.1 it is clear that the former is certainly true and and thus by implication so is the latter. However more is required than this from the aid for it to have remedial potential. We require that from the CASTT we can see from the plots when sounds, words and phrases are segmentally and suprasegmentally similar (e.g. a same-speech pair) and when sounds, words and phrases are segmentally and suprasegmentally different (e.g a different-speech pair).

We can see from curve (b) in figure 6.1 a sizeable proportion of the pairs of plots

the participants selected as being most similar were from same speech pairs. Curve (b) in fig 6.1 (b) is the cumulative graph of the number of same speech pairs correctly identified by at least z participants, where z = is 20 to 31. From this graph it can be seen that 141 out of the 720 (i.e. 19.6 %) same speech pairs were correctly identified by all of the participants. Thus 77.9 % (141/181) of the pairs of plots which were identified as being the most similar by all the participants were in fact the displays of same-speech pairs. 295 (40.9 %) of the same-speech pairs were correctly identified by 28 or more of the participants. Thus 78.5 % (295/376) of the pairs of plots which were identified as most similar by 90.3% or more of the participants were from same-speech pairs.

It can be seen from curve (a) and curve (b) in fig 6.1 that if a certain pair of plots is considered to be the most similar by a large portion of the participants (80% or more) then between 70% and 75% of the time it will in fact be the pair of displays for a same-speech pair. This is very promising since for the CASTT to be useful the displays should only be perceived as similar if they are from utterances that are phonetically and suprasegmentally similar.

We have argued that it must be accepted that the CASTT will have remedial potential of less than 100% success in distinguishing target and error utterances. However we have not established what this percentage must be. Obviously whilst it will not be 100% it must still be a high proportion of the participants. Looking at the gradient of the curve (b) in figure 6.1 it can be seen the gradient changes noticeably near 28 participants. For this reason it was decided to set the percentage of participants which defines remedial potential at the 28 out of 31 level. A second level at 25 out of 31 participants was also included as a less demanding goal.

The first level was called the 90+ % level. At this level if 28 or more participants (90.3 % or more) distinguish the plots of target and error utterances for a particular target/error combination then for the elementary error, which is exemplified by the combination, the CASTT was said to have remedial potential at the 90+ % level. In order for this to be achieved the same-speech pair in each of the four plot-sets associated with each elementary error must each be correctly identified by 28 (90.3 %) or more of the participants.

The second level was called the 80+ % level. This level was defined in exactly the same manner as the 90+ % level but it was for 25 (80.6 %) or more of the participants rather than 28 or more.

The target/error combinations of NZ-SL2 exemplify the same elementary errors as their respective counterparts in NZ-SL1. However the NZ-SL2 combinations only comprise isolated phones whereas those in NZ-SL1 comprise words or syllables. Whilst NZ-SL2 could be used to test the visual displays of the modules with time-plots (pitch contours, loudness contours or spectral content response) NZ-SL1 cannot be used to test the modules with current-value-plots (fricative content response, vocal tract shape estimation or Lissajous figure). If there is an elementary error for which the CASTT has remedial potential with one of the current-value-plot display types but not one of the time-plot display types, it means we can only seek to remediate the speech error in isolated phones. We have no facility for checking the carry-over of the correctly pronounced phone into syllables, words and in continuous speech. In this situation the CASTT has remedial potential in a limited form only.

The VDT tested the visible speech aid, the CASTT, both for its ability to separate

| Female | | Male | |
|-------------|-------------|-------------|-------------|
| 90+ % level | 80+ % level | 90+ % level | 80+ % level |
| EE-1 | EE-1 | EE-1 | EE-1 |
| EE-4 | EE-4 | | EE-4 |
| | EE-5 | | |
| | EE-7 | | |
| EE-10 | EE-10 | EE-10 | EE-10 |
| | EE-11 | EE-11 | EE-11 |
| | EE-12 | | |
| EE-13 | EE-13 | | EE-13 |
| EE-15 | EE-15 | EE-15 | EE-15 |
| EE-16 | EE-16 | | |
| | EE-18 | | |
| EE-21 | EE-21 | EE-21 | EE-21 |
| EE-23 | EE-23 | | |
| EE-26 | EE-26 | | |
| | EE-28 | | |

Table 6.6. The elementary errors, exemplified by the target/error utterances in NZ-SL1 for which CASTT has remedial potential at the 90+ % and 80+ % level. The data is given for both female and male speech.

different-speech pairs and for its ability to display same-speech pairs as less different than different-speech pairs. Thus we can only talk about remedial potential rather than remedial efficacy because the VDT does not establish that a visible speech aid can be used to remediate speech. That can only be established by conducting a clinical trial with the speech aid in a therapy situation.

6.4.5 The Results Of The VDT

Table 6.6 shows the CASTT's remedial potential for the elementary errors exemplified by the target/error combinations in NZ-SL1 at the 90+ % level and the 80+ % level from the plots obtained from female and male speech. The results were obtained from the spectral content, pitch contour and loudness contour display types. Table 6.7 shows the CASTT remedial potential for the elementary errors exemplified by the target/error combinations in NZ-SL2 at the 90+ % level and the 80+ % level from the plots obtained from female and male speech. These results were obtained from the responses to the fricative content response, vocal tract shape estimation and Lissajous figure display types.

The results in table 6.6 and table 6.7 are well above what one would get if random guessing was employed to make the "most similar" decisions. The probability of successfully identifying all four same speech pairs, if random guessing was used to make the decisions, at the 90+ % level was 4.85855×10^{-42} and at the 80+ % level was 2.36974×10^{-29} . These probabilities were calculated assuming the binomial distribution, using (B.1) in appendices B.2. We already knew that the results would be well above random guessing because of the large proportion of the same-speech pairs which were being identified correctly by a significant number of the participants.

The results in table 6.6 and table 6.7 were certainly far more promising than those obtained from the preliminary VDT, see table 6.4. For female speech the CASTT had

| Female | | Male | |
|---------------|-------------|-------------|-------------|
| 90+ % level | 80+ % level | 90+ % level | 80+ % level |
| EE-9 EE-10 | EE-8 | EE-6 | EE-6 |
| | EE-9 | EE-8 | EE-8 |
| | EE-10 | EE-10 | EE-10 |
| | | EE-12 | EE-12 |
| | | EE-13 | EE-13 |
| | EE-14 | | |
| | | EE-15 | EE-15 |
| EE-17 | EE-17 | EE-17 | EE-17 |

Table 6.7. The elementary errors, exemplified by the target/error utterances in NZ-SL2 for which the CASTT has remedial potential at the 90+ % and 80+ % level. The data is given for both female and male speech.

remedial potential for 15 of the 29 elementary errors exemplified by the target/error utterances in NZ-SL1 and 5 of the 12 elementary errors exemplified by the target/error utterances in NZ-SL2, at the 80+ % level. For male speech the CASTT had remedial potential for 7 out of the 29 elementary errors exemplified in NZ-SL1 and 7 out of the 12 elementary errors exemplified in NZ-SL2 at the 80+ % level.

There were still a large number of elementary errors, exemplified by utterances from either speaker, for which the CASTT had no remedial potential. These errors are listed in table 6.8. It can be seen in this table there were ten elementary errors exemplified by utterances made by the female speaker for which the CASTT had no remedial potential. There were a further four errors, EE-8, EE-9, EE-14 and EE-17 for which the CASTT had remedial potential from one of the current-value-plot display types but not the time-plot display types. Thus the remedial potential was in a limited form. There were eighteen elementary errors exemplified by the utterances of the male speaker in which the CASTT has no remedial potential (see table 6.8). In addition there were a further four errors, EE-6, EE-8, EE-12 and EE-17 in which the CASTT had limited remedial potential. For the nine elementary errors EE-2, EE-3, EE-19, EE-20, EE-22, EE-24, EE-25, EE-27 and EE-29, exemplified by the utterances of either speaker, the CASTT had no remedial potential in any form at all.

In several instances some of the elementary errors listed in table 6.8 had three out of the four same-speech pairs successfully identified at the 80+ %. These elementary errors are marked in the table with a †(e.g. EE2†). The fourth same-speech in all the above cases was correctly identified by well over the majority of the participants. Clearly the CASTT almost has remedial potential for these errors. If the visual cues from which people made their “most similar” decisions were somehow emphasized then perhaps the CASTT would have remedial potential for these errors.

6.4.5.1 The Remedial Potential Of Each Of The Visual Display Types

Figure 6.2 gives the number of participants who correctly identified the same-speech pair in each of the plot-sets of the loudness contours. The loudness contours were of utterances by the female speaker. Figure 6.3 is similar to figure 6.2 but the loudness contours were of utterances by the male speaker. The results in fig 6.2 and 6.3 are

| Female | | Male | |
|--------|--------|--------|--------|
| NZ-SL1 | NZ-SL2 | NZ-SL1 | NZ-SL2 |
| EE-2 | EE-2 | EE-2† | EE-2 |
| EE-3† | EE-6 | EE-3† | EE-9 |
| EE-6† | EE-12 | EE-5 | EE-14 |
| EE-19 | EE-13 | EE-7† | EE-16 |
| EE-20 | EE-15 | EE-9 | EE-19 |
| EE-22† | EE-16 | EE-14 | |
| EE-24 | EE-19† | EE-16 | |
| EE-25 | | EE-18 | |
| EE-27 | | EE-19† | |
| EE-29 | | EE-20 | |
| | | EE-22† | |
| | | EE-23 | |
| | | EE-24 | |
| | | EE-25 | |
| | | EE-26† | |
| | | EE-27 | |
| | | EE-28 | |
| | | EE-29 | |

Table 6.8. The elementary errors for which the CASTT has no remedial potential. The elementary errors marked with † had three of the four same speech pairs successfully identified at the 80 % level.

arranged according to elementary error. There are four columns associated with each elementary error. The number of participants who correctly identified the same-speech pair X_1X_2 in the plot-set $X_1X_2Y_1$ is given by the height of the first of the four columns. The height of the second column is the number of participants who correctly identified the same-speech pair X_1X_2 in the plot-set $X_1X_2Y_2$. The heights of the third and fourth columns are the numbers of participants who correctly identified the same-speech pair Y_1Y_2 in the plot-sets $X_1Y_1Y_2$ and $X_2Y_1Y_2$ respectively.

The elementary errors in which the loudness contour display type has remedial potential at the 90+% level are indicated in figures 6.2 and 6.3 by all four columns associated with the error being coloured black (eg. EE-10 in figure 6.2). The elementary errors in which the loudness contour display type has remedial potential at the 80+% level are indicated by all four columns associated with the error being coloured dark grey (eg. EE-1 in figure 6.2). The elementary errors in which the loudness contour display type has no remedial potential are indicated by the light grey columns (eg. EE-2 in figure 6.2). The reader is reminded that the target/error combinations which exemplify each of the elementary errors are given in tables 6.1, 6.2 and 6.3.

Figures 6.4, 6.6, 6.10, 6.8 and 6.12 give the number of participants who correctly identified the same-speech pair in each of the plot-sets from the pitch contours, the spectral plots, the vocal tract shapes, the fricative response displays and the lissajous figures respectively. All the visual displays in these figures were from utterances by the female speaker. The layout of these graphs is exactly the same as those in figures 6.2 and 6.3, which were described in detail above. Figures 6.5, 6.7, 6.11, 6.9 and 6.13 are similar to the above mentioned figures but the visual display types were from utterances of the male speaker.

Some of the results in figures 6.2 to 6.13 were unexpected. For example the VDT

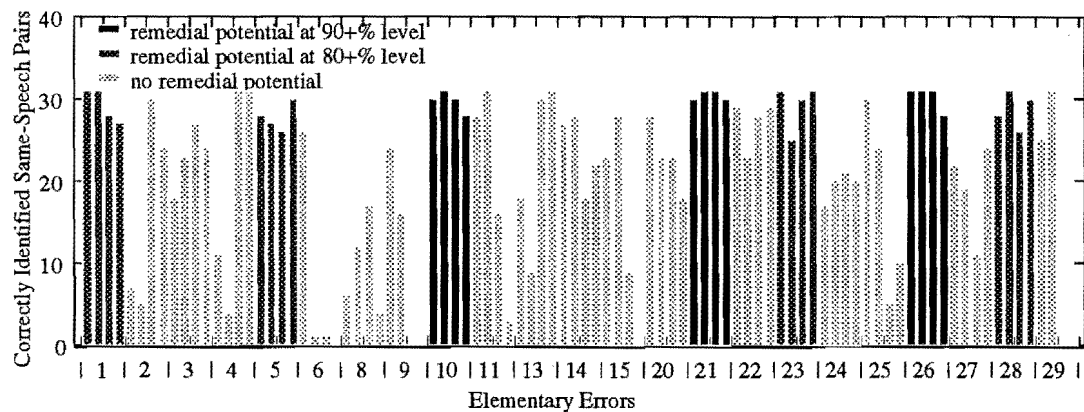


Figure 6.2. The numbers of participants who correctly identified the same-speech pairs in each of the plot-sets of the loudness contours. The loudness contours were of utterances by the **female** speaker. The elementary errors, for which the loudness contours of utterances exemplified by female utterances had remedial potential at the 90+% level and at the 80+% level, are indicated by the columns associated with the errors being colored black and dark grey respectively.

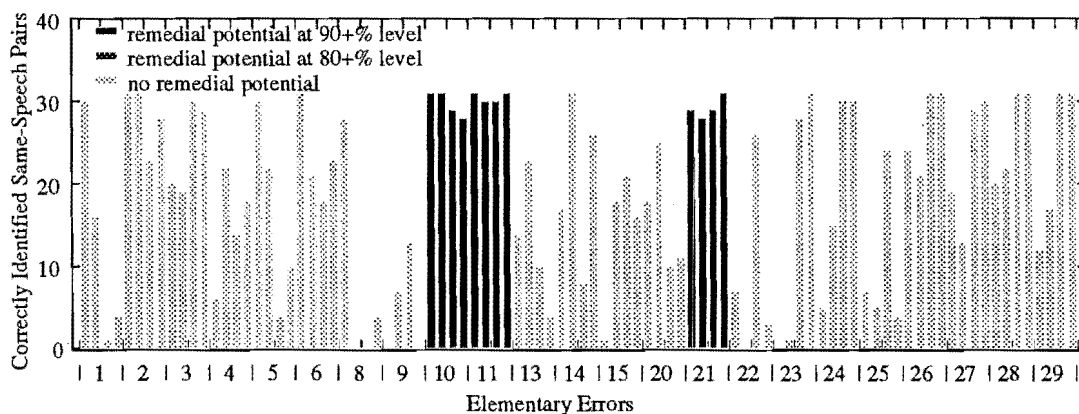


Figure 6.3. The numbers of participants who correctly identified the same-speech pairs in each of the plot-sets of the loudness contours. The loudness contours were of utterances by the **male** speaker. The elementary errors, for which the loudness contours of utterances exemplified by male utterances had remedial potential at the 90+% level and at the 80+% level, are indicated by the columns associated with the errors being colored black and dark grey respectively.

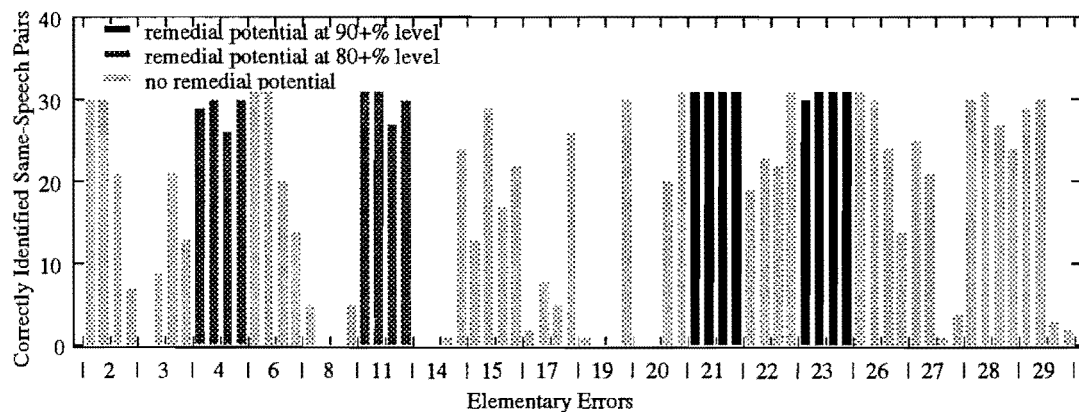


Figure 6.4. The numbers of participants who correctly identified the same-speech pairs in each of the plot-sets of the pitch contours. The pitch contours were of utterances by the **female** speaker. The elementary errors, for which the pitch contours of utterances exemplified by female utterances had remedial potential at the 90+% level and at the 80+% level, are indicated by the columns associated with the errors being colored black and dark grey respectively.

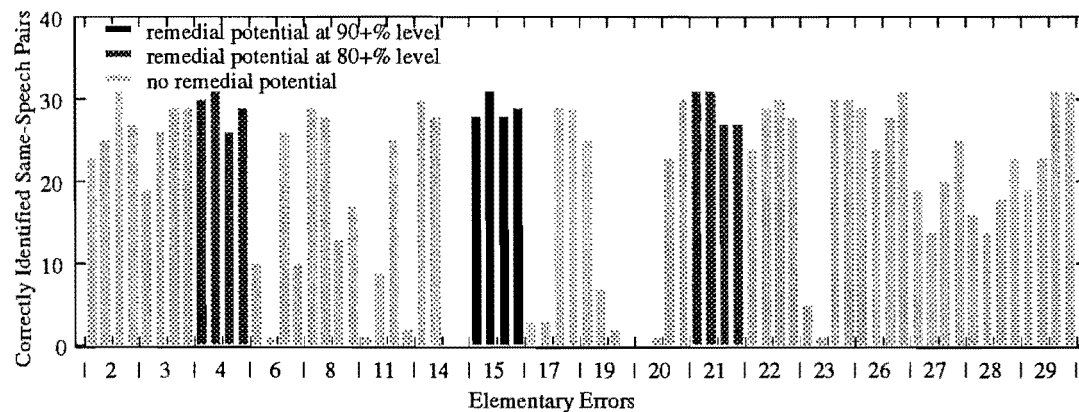


Figure 6.5. The numbers of participants who correctly identified the same-speech pairs in each of the plot-sets of the pitch contours. The pitch contours were of utterances by the **male** speaker. The elementary errors, for which the pitch contours of utterances exemplified by male utterances had remedial potential at the 90+% level and at the 80+% level, are indicated by the columns associated with the errors being colored black and dark grey respectively.

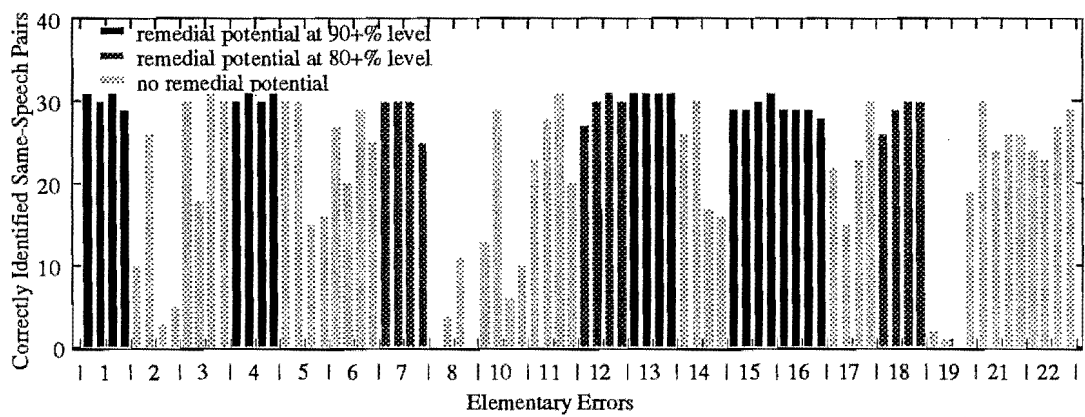


Figure 6.6. The numbers of participants who correctly identified the same-speech pairs in each of the plot-sets of the spectral content plots. The spectral content were of utterances by the female speaker. The elementary errors, for which the spectral content plots of utterances exemplified by female utterances had remedial potential at the 90+% level and at the 80+% level, are indicated by the columns associated with the errors being colored black and dark grey respectively.

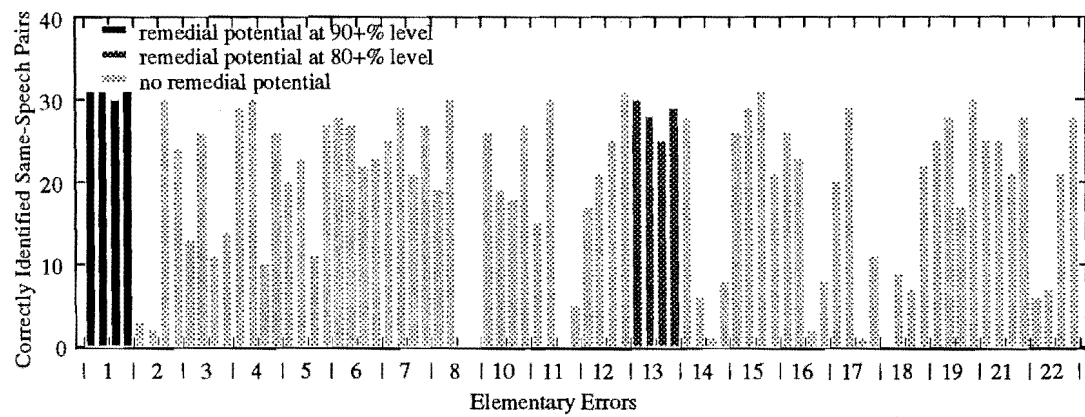


Figure 6.7. The numbers of participants who correctly identified the same-speech pairs in each of the plot-sets of the spectral content plots. The spectral content plots were of utterances by the male speaker. The elementary errors, for which the spectral content plots of utterances exemplified by male utterances had remedial potential at the 90+% level and at the 80+% level, are indicated by the columns associated with the errors being colored black and dark grey respectively.

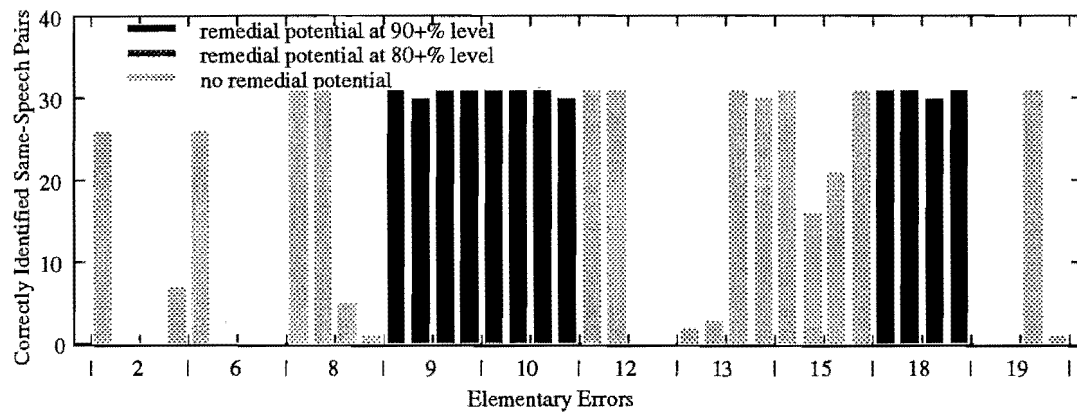


Figure 6.8. The numbers of participants who correctly identified the same-speech pairs in each of the plot-sets of the fricative content plots. The fricative content plots were of utterances by the female speaker. The elementary errors, for which the fricative content plots of utterances exemplified by female utterances had remedial potential at the 90+% level and at the 80+% level, are indicated by the columns associated with the errors being colored black and dark grey respectively.

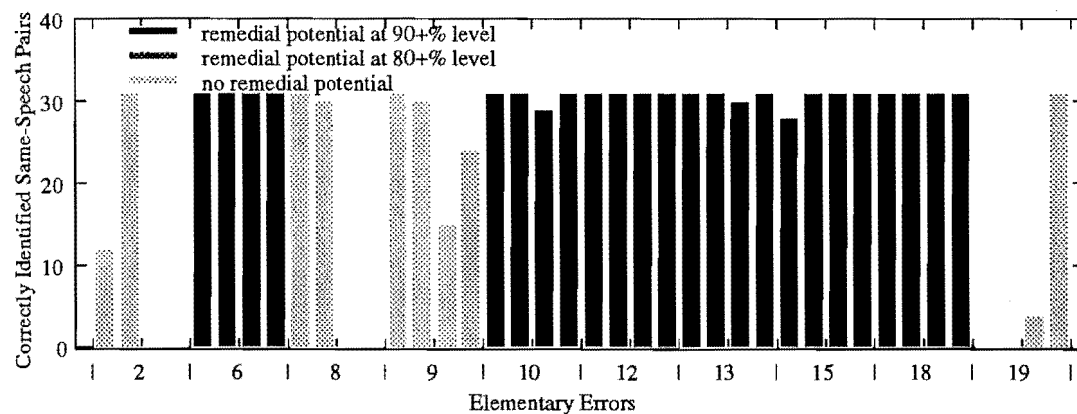


Figure 6.9. The numbers of participants who correctly identified the same-speech pairs in each of the plot-sets of the fricative content plots. The fricative content plots were of utterances by the male speaker. The elementary errors, for which the fricative content plots of utterances exemplified by male utterances had remedial potential at the 90+% level and at the 80+% level, are indicated by the columns associated with the errors being colored black and dark grey respectively.

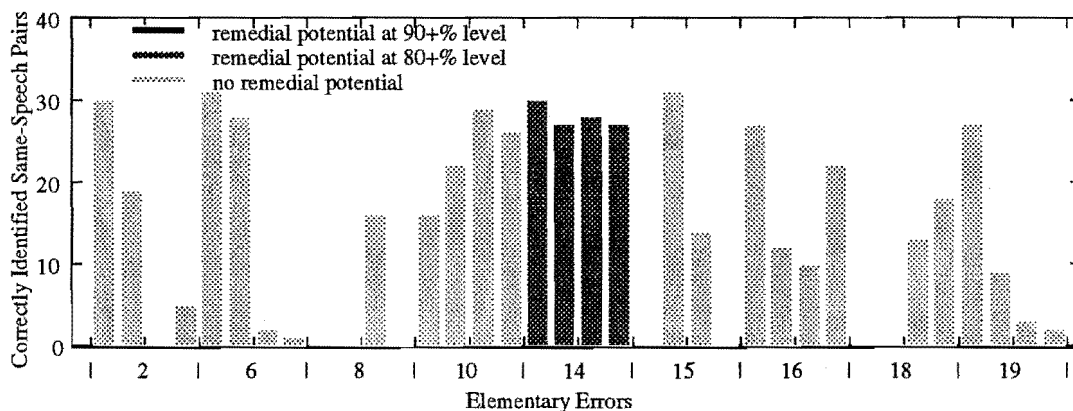


Figure 6.10. The numbers of participants who correctly identified the same-speech pairs in each of the plot-sets of the vocal tract shape reconstructions. The vocal tract shape reconstructions were of utterances by the **female** speaker. The elementary errors, for which the vocal tract shape reconstructions of utterances exemplified by female utterances had remedial potential at the 90+% level and at the 80+% level, are indicated by the columns associated with the errors being colored black and dark grey respectively.

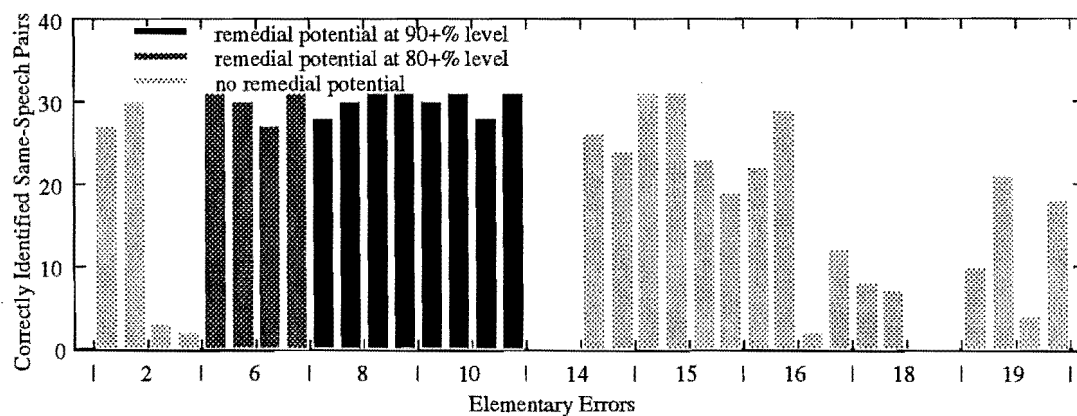


Figure 6.11. The numbers of participants who correctly identified the same-speech pairs in each of the plot-sets of the vocal tract shape reconstructions. The vocal tract shape reconstructions were of utterances by the **male** speaker. The elementary errors, for which the vocal tract shape reconstructions of utterances exemplified by male utterances had remedial potential at the 90+% level and at the 80+% level, are indicated by the columns associated with the errors being colored black and dark grey respectively.

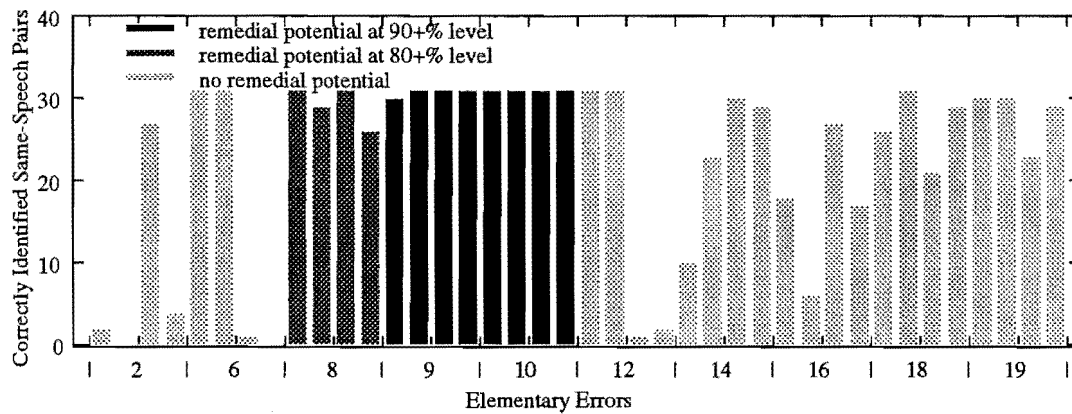


Figure 6.12. The numbers of participants who correctly identified the same-speech pairs in each of the plot-sets of the Lissajous figures. The Lissajous figures were of utterances by the **female** speaker. The elementary errors, for which the Lissajous figures of utterances exemplified by female utterances had remedial potential at the 90+% level and at the 80+% level, are indicated by the columns associated with the errors being colored black and dark grey respectively.

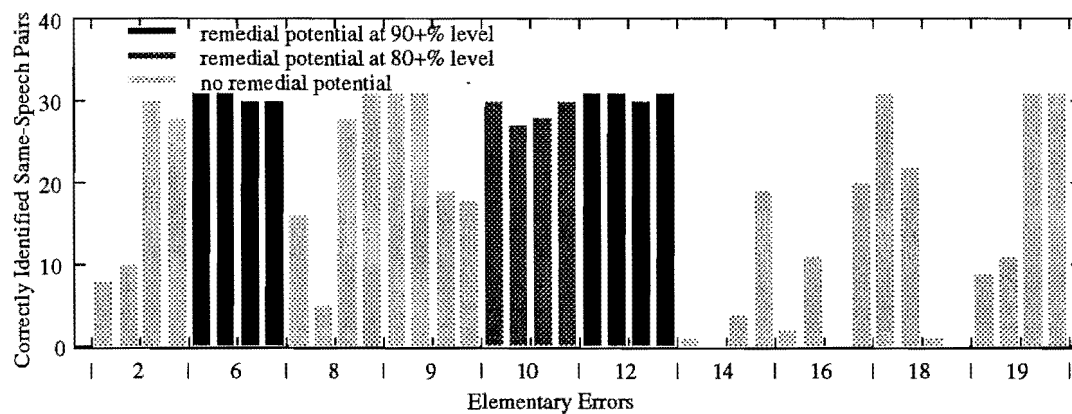


Figure 6.13. The numbers of participants who correctly identified the same-speech pairs in each of the plot-sets of the Lissajous figures. The Lissajous figures were of utterances by the **male** speaker. The elementary errors, for which the Lissajous figures of utterances exemplified by male utterances had remedial potential at the 90+% level and at the 80+% level, are indicated by the columns associated with the errors being colored black and dark grey respectively.

indicated the CASTT had no remedial potential in EE-24 (errors in excessive loudness) and EE-25 (errors in insufficient loudness). We expect the differences in the average vocal intensity to be observable by the differences in the average height of the loudness contours. We know from clinical experience that this does happen (see sec 5.2.1), and further more it happened in the loudness contours in the VDT. Figure 6.14 gives the loudness contours of the phrase " Grape juice and water mix well" spoken normally and spoken softly. The bottom two contours are clearly much lower in height than the upper two contours. In spite of this, each of the same-speech pairs in the four plot-sets associated with EE-25 were not correctly identified by at least 80 % of the participants. Thus the loudness contour display type had no remedial potential for that elementary error. Clearly there is still some design fault in the VDT!

The VDT was designed so that the participants had no knowledge of the utterances the plots were from. This was done intentionally in order that the participants' ability to interpret the plots was not conflated with CASTT's capacity to display speech errors. This structure meant some very strong results could be obtained on what features in the displays the participants' considered to be important. However there were two fundamental problems with the VDT.

The VDT was structured in such a way that it made an implicit assumption. It assumed that if a display type had remedial potential for an elementary error then for the displays of the target and error utterances which exemplified the error the most obvious difference between the displays was caused by the intended acoustic difference between the two utterances. Something as important as this cannot be assumed; it must be tested.

Since the participants did not know what the target and error utterances were they did not know what features they should be looking for evidence of. This was an intended aspect of the VDT, but it gave the impression the CASTT was less useful than clinical experience suggests it could be. A cursory inspection of the displays in the VDT revealed that there were instances in which there were differences between the displays of the target and error utterances which were directly related to the intended acoustic differences but the participants did not correctly identify all the four same-speech pairs. The participants were distracted by some other visual feature which was usually unrelated to the acoustic event of interest. For example in fig 6.14 the loudness contours of one of the error utterances had fewer segments than the other contour. This was because the phrase was uttered exceedingly softly and no loudness contours could be calculated for some parts of the phrase.

In a speech therapy situation the therapist would point out to the client which features in a display to be aware of. A new test has been designed to test the number of elementary errors the CASTT has remedial potential for when the user has knowledge of what features to focus on in the displays from the target and error utterances in the VDT.

6.5 THE VISUAL DISPLAY TEST PART II

In the new visual test all the displays of the VDT were re-examined. The new test will be referred to as VDT part II. The previously discussed VDT will now be called VDT part I. In VDT part II evidence of the intended acoustic differences between the target and error utterances was looked for in the displays of the utterances. The displays

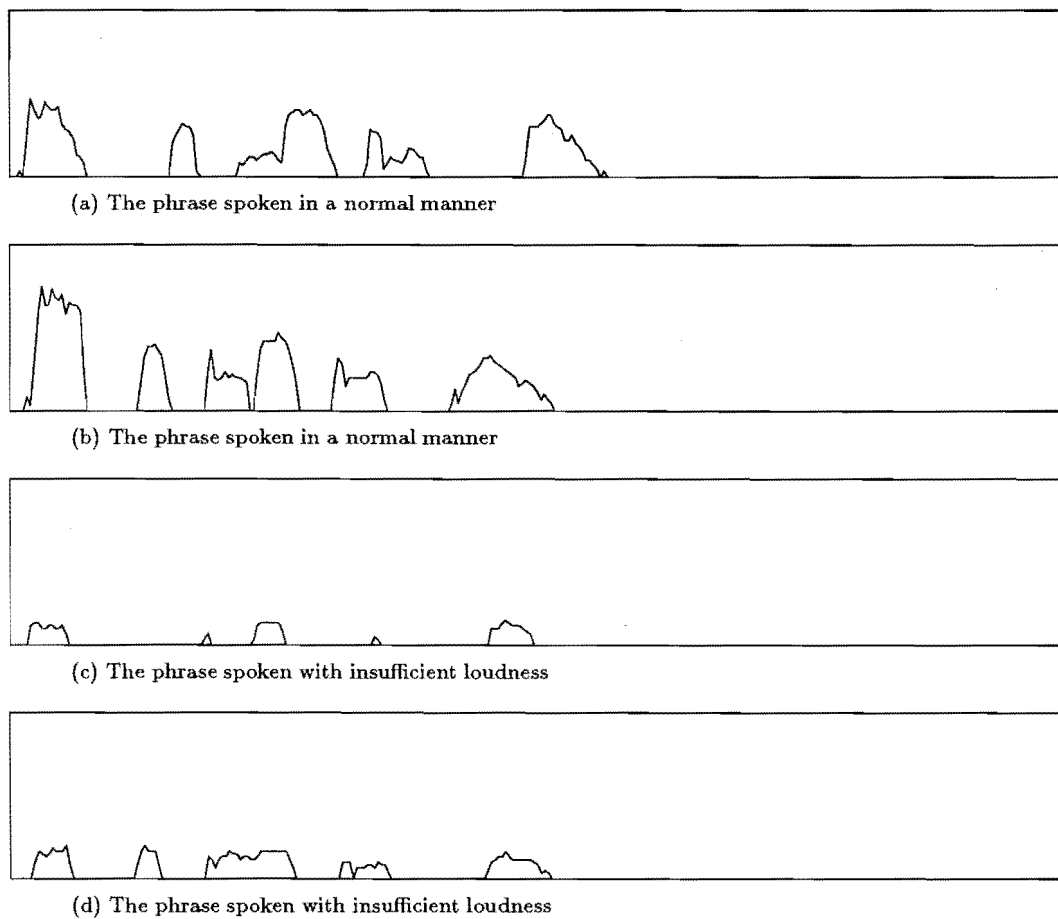


Figure 6.14. The contours of the target and error utterances, spoken by the female speaker, which exemplify EE-25. The contours of the target phrases "Grape juice and water mix well" spoken in a normal voice are given in the top two displays. The contours of the error, the above mentioned phrase spoken with insufficient loudness, are given on the bottom two displays.

| | | ART. INT. | | | | | UN./V. | | | NAS. | | ART. SUB. | | | SUPR. | | | SP. TIMB. | | | | | | | | | |
|--------|-----------|-----------|---|---|---|---|--------|---|---|------|----|-----------|----|----|-------|----|----|-----------|----|----|----|----|----|----|--|--|--|
| TEST | UTTERANCE | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 9 | 10 | 11 | 13 | 14 | 15 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | | | |
| VDT II | Female | ✓ | | ✓ | ✓ | ✓ | | | | ✓ | | | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | |
| VDT II | Male | ✓ | ✓ | ✓ | ✓ | | | | | ✓ | ✓ | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | | | |
| VDT I | Female | ✓ | | | | ✓ | | | | ✓ | | | | | | ✓ | | ✓ | | | ✓ | | ✓ | | | | |
| VDT I | Male | | | | | | | | | ✓ | ✓ | | | | | ✓ | | | | | | | | | | | |

Table 6.9. The Elementary Errors for which the loudness contours had remedial potential in VDT part I and II are indicated by a tick.

were judged according to four criteria. It was required that the target displays be more similar than the displays of the target/error different speech pairs. Secondly the error displays had to be more similar than the displays of the error/target different speech pairs. The target and error utterances had to differ intentionally in one main feature of speech, which might be an allophone or an aspect of speech timbre, etc. Thus the third requirement was that the presence or absence of the speech feature of interest must be obvious from the visual display. Finally it was required that the visual difference between the target and error displays be related to the intended acoustic difference between the target and error utterances. If all four conditions were met then it was said that the display type had remedial potential for the elementary error the target and error utterances exemplified. Note that the criteria for remedial potential in VDT part II are quite different from those in VDT part I (see sec 6.4.4).

The displays in this test were exactly the same as those used in the VDT part I, as were the elementary errors for which each display type was tested for remedial potential. In this test the displays of the VDT part I were carefully re-examined by one person - myself. The results of the re-examination are discussed in sections 6.5.1 and 6.5.2. Section 6.5.1 discusses the results for the time-plots and sec 6.5.2 discusses the results for the current-value-plots. The VDT part II revealed that all the time-plot display types were quite powerful speech tools, but the current-value-plot display types were not. This is contrary to the findings in VDT part I and extends the remedial potential of the CASTT significantly.

6.5.1 The Time-Plots

6.5.1.1 The Displays Of Loudness Contours

Table 6.9 gives the elementary errors for which the loudness contours had remedial potential in VDT part II. It is divided into the six speech error sets of the Elementary Error list, given in tables 6.1 and 6.2. The elementary errors exemplified by the female and male utterances for which the loudness contours have remedial potential have been listed separately. For comparison the elementary errors for which the contours had remedial potential for in VDT part I have also been included. Each of the elementary errors for which the loudness contours have remedial potential are indicated by a tick. It can be seen immediately from table 6.9 that the loudness contours have remedial potential for many more elementary errors in VDT part II than in VDT part I.

The following is a discussion about the contours of the utterances which exemplified the elementary errors for which the loudness contours had remedial potential. Utterances by the female speaker will henceforth be called female utterances. Similarly utterances by the male speaker will be called male utterances. The loudness contours

examined in VDT part II will be discussed with reference to the speech error sets in the Elementary Error list.

The Articulation Intensity set comprised errors which were exemplified by target and error utterances which intentionally differed in the shapes of the envelopes of the time domain wave forms. For this reason we expected to see differences in the shapes of the loudness contours. All the target and error utterance pairs differed in one phoneme. The utterances which exemplified EE-3 and EE-5 differed in the initial phoneme, those which exemplified EE-2 differed in the medial phoneme and the utterances which exemplified EE-1 and EE-5 differed in the final phoneme. These facts gave an indication where one should look in the contours to distinguish between the target and error utterances. It was also necessary to take into account the co-articulation effects due to the differing phonetic contexts in the target and error utterances; for example the vowel [u] is longer in "boo" than in "boot".

In the VDT part I, the loudness contours, from female utterances, had remedial potential for both of EE-1 and EE-5 (see fig 6.2). The loudness contours of the target and error utterances which exemplified those errors were obviously distinguishable by shape, duration and the heights of the contour. So much so that the participants in VDT part I were able to correctly identify the same-speech pairs without being told what feature(s) to focus on. The contours of the target and error utterances which exemplify EE-1 and EE-5 also fulfill the four requirements for remedial potential in VDT part II. In fig 6.15 we can see both the contours of "boo" look similar as do both the contours of "boot". In addition we can see the contours of "boo" are much wider than those of "boot", due to the longer duration of the vowel [u] in "boo".

The VDT part II, in fact revealed that the loudness contours had remedial potential for most of the elementary errors in the Articulation Intensity set, regardless of the speaker. The loudness contours had remedial potential for EE-1, EE-3, EE-4 and EE-5, exemplified by the female utterances and EE-1, EE-2, EE-3 and EE-4, exemplified by the male utterances (see table 6.9).

The contours of the male utterances which exemplified EE-2 were distinctly distinguishable by shape. In fig 6.16 we can see there is a dip in the middle of the contours of "allay", and a peak in the middle of the contours of "away". In VDT part I, these loudness contours did not have remedial potential for EE-2. However we can see from fig 6.3 that they almost did. The same-speech pairs were correctly identified by at least 80 % of the participants in three of the four plot-sets. In the fourth plot-set, which consisted of the left contour of "allay" and the two contours of "away", 23 out of the 31 (74 %) participants correctly identified the same-speech pair. If one is aware that the middle section of the contour was the feature which should be focused on ("allay" and "away" differ in the medial phoneme) then it can be seen that the contours in fig 6.16 fulfil the four requirements of remedial potential in VDT part II.

The loudness contours of the target and error female utterances which exemplify EE-3 and EE-4, were also distinguishable by shape; as were the contours of the male utterances which exemplified EE-1, EE-3 and EE-4. However the ability to distinguish between the target and error contours was not obvious at first glance.

The target/error utterances which exemplify EE-1 both have the phoneme [b] in the initial position. When this plosive is pronounced it can be fully or partially released. All [b] sounds in the initial position are fully released, however they can differ in the strength of the release and hence in intensity. The target male utterances both

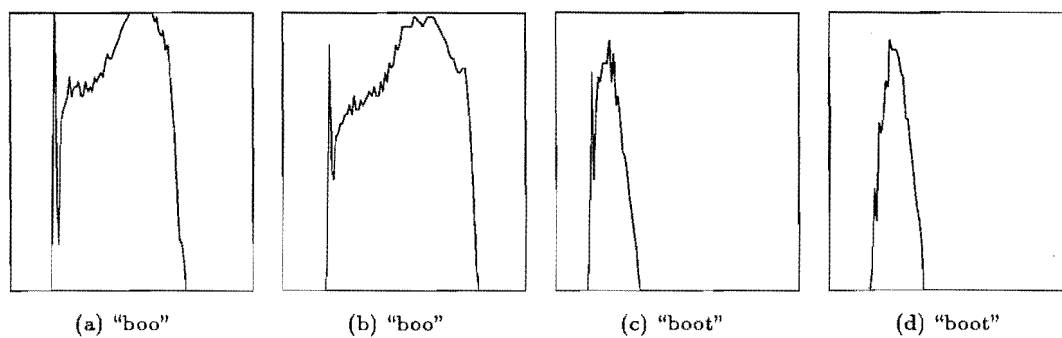


Figure 6.15. The contours of the target and error female utterances "boo" and "boot", which exemplify EE-1. The contours of the target words are given in (a) and (b) and those of the error words given in (c) and (d).

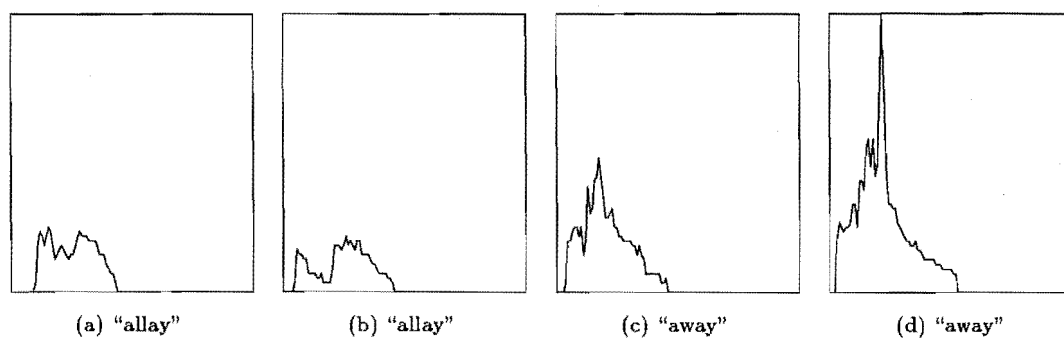


Figure 6.16. The contours of the target and error male utterances "allay" and "away", which exemplify EE-2. The contours of the target words are given in (a) and (b) and those of the error words given in (c) and (d).

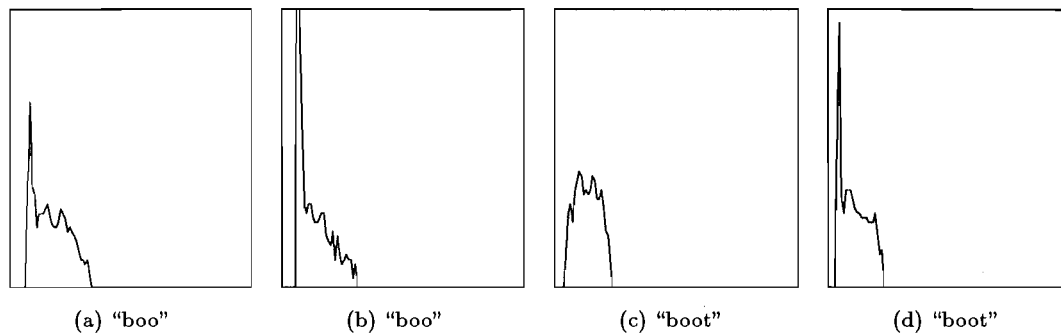


Figure 6.17. The contours of the target and error male utterances “boo” and “boot”, which exemplify EE-1. The contours of the target words are given in (a) and (b) and those of the error words given in (c) and (d).

contained the strongly released [b], as seen by the spike in the contours of “boo” in fig 6.17. One of the error utterances of “boot” contained a strongly released [b], as shown by the spike in the contour of “boot” in fig 6.17 (d). However the [b] in the other utterance of “boot” was only weakly released. The distinguishing feature between “boo” and “boot”, for the loudness contours, is the long [u] in “boo”. Thus ignoring any effects at the start of the contours due to the strength of the released plosive, it can be seen in fig 6.17 that both contours of “boo” look similar as do both contours of “boot” the contours of “boo” being slightly wider than those of “boot” and falling more gradually. Hence the contours of the target and error male utterances which exemplify EE-2 fulfil the requirement for remedial potential in VDT part II.

It is interesting to note that not all the [b]’s were strongly released in the female utterances of “boot”, as can be seen in fig 6.15. However since the contours of “boo” are so much wider than those of “boot” the absence of the spike in the contour “boot” in fig 6.15(d) is not very noticeable.

The effect of fully releasing a plosive, or partially releasing it, also caused a distracting effect on the contours of the target and error utterances which exemplified EE-4. The initial [p] in “poppa” was fully aspirated in only one of the female and male utterances. The initial [p] in “pop” was fully aspirated in both the male utterances but only in one of the female utterances. Figure 6.18 shows the contours of “poppa” and “pop” spoken by the female speaker. Ignoring the effect on the contours due to the initial [p], it can be seen in fig 6.18 that the contour segments of the sound [pɒ] in “pop” are wider than the contour segments of the sound [pɒ] in “poppa”. This corresponds to the vowel [ɒ] being more stressed in “pop” than in “poppa”. It can also be seen in fig 6.18 that the contours of the syllable [pʌ] in “poppa” are longer than the contours of the final [p] in “pop”. In addition the final [p] in “pop”, in both utterances, was fully aspirated but in both utterances of “poppa” it was only partially aspirated.

The difference between the more stressed and less stressed [pɒ] syllable in “pop” and “poppa” could also be seen in the contours of the male utterances, along with the differences in duration between the second [p] in “pop” and the final syllable [pʌ] in “poppa”. The second [p]’s in both “poppa” and “pop”, when spoken by the male speaker, were all fully aspirated. A figure with these contours was not included because they are not controversial. Thus for the reasons just outlined, the contours of the

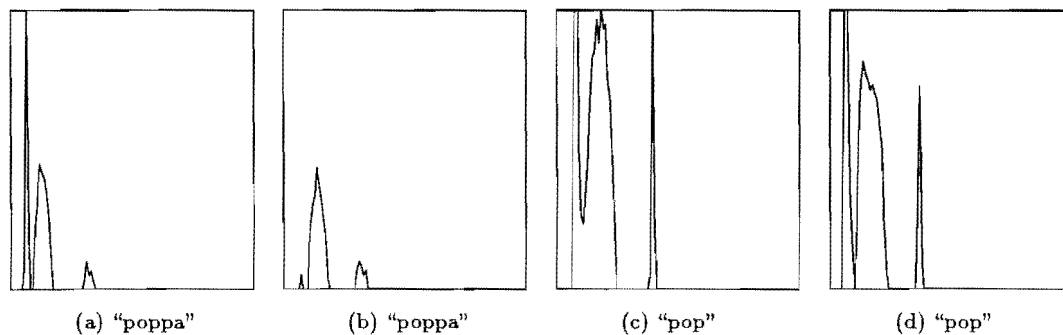


Figure 6.18. The contours of the target and error female utterances “poppa” and “pop”, which exemplify EE-4. The contours of the target words are given in (a) and (b) and those of the error words given in (c) and (d).

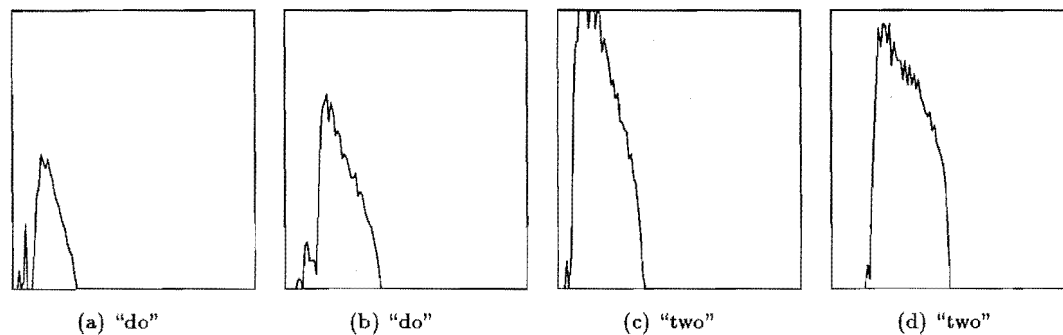


Figure 6.19. The contours of the target and error female utterances “do” and “two”, which exemplify EE-3. The contours of the target words are given in (a) and (b) and those of the error words given in (c) and (d).

utterances, spoken by either speaker, which exemplify EE-4 fulfil the requirement for remedial potential in VDT part II.

To distinguish between the contours of “two” and “do”, the target and error utterances which exemplified EE-3, it was necessary to consider the effects of the initial phonemes in the two words and the related co-articulation effects. The allophonic realizations of the initial phonemes were the unvoiced alveolar plosive [t] and its voiced pair [d]. The voice onset time is greater for an unvoiced plosive than for a voiced one. It was evidence of this that we looked for in the contours. Figure 6.19 gives the contours of “two” and “do” spoken by the female speaker. It can be seen that there is a greater length between the start of the contour segments associated with [d] and that associated with [u] than the starts of the contour segments associated with [t] and [u]. This suggests the voice onset time was greater in the utterance of “two” than in “do”. Thus the intended acoustic difference between “two” and “do” is observable in the loudness contours. The same sort of features in the loudness contours of “two” and “do” spoken by the male speaker could also be seen. Hence the contour of the target and error utterance, from either speaker, which exemplify EE-3 fulfil the four requirements for remedial potential of VDT part II.

There was no difference in the results of the VDT part I and VDT part II in the

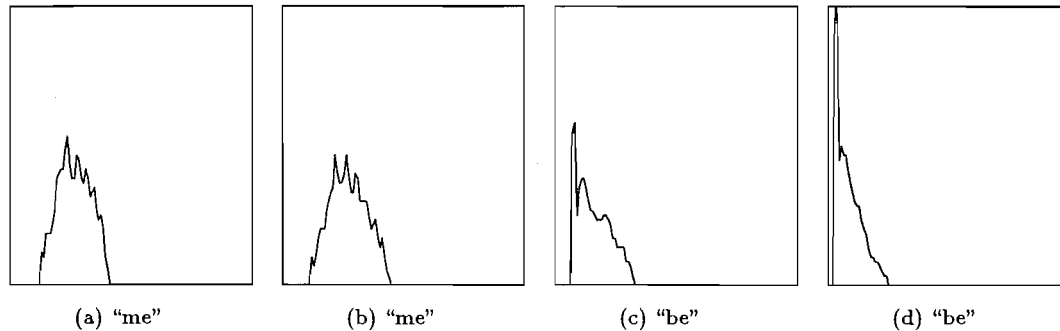


Figure 6.20. The contours of the target and error male utterances of "me" and "be", which exemplify EE-10. The contours of the target words are given in (a) and (b) and those of the error words given in (c) and (d).

number of elementary errors in the Unvoiced/Voiced set and the Nasal set for which the loudness contours had remedial potential (see table 6.9). In VDT part II none of the target and error utterances, from either speaker, which exemplified the elementary errors in the Voiced/Unvoiced set were distinguishable from their loudness contours. However most of the target and error utterances which exemplified the elementary errors in the Nasal Quality set were distinguishable by shape of the contours.

The loudness contours of the target and error male utterances which exemplified EE-10 and EE-11 were very distinctive; as were the contours from the female utterances which exemplified EE-10. This is supported by the fact that loudness contours had a remedial potential at the 90 % level in the VDT part I, for EE-10, exemplified by utterances from both speakers (see figs. 6.2 and 6.3) and for EE-11 exemplified by utterances from the male speaker (see 6.3). Thus the differences between these target and error contours were distinct enough that no knowledge of the acoustic features which distinguished the target and error utterances was necessary.

Figure 6.20 is an example of one of the sets of contours in the Nasal set. It gives the contours of "me" and "be", the utterances which exemplify EE-10, spoken by the male speaker. Without having any knowledge of what the contours are of, it can be seen that the two contours of "be" (fig 6.20 (c) and (d)) both have spikes in them whereas the two contours of "me" (fig 6.20 (a) and (b)) are more hoop-like. The difference between the utterance "me" and "be" was in the initial phonemes and the related co-articulation effects. The spike in the contours of "be" is due to the fact the [b] was fully released. In the contours of "me" the highest region of the contours is wiggly. This is possibly due to the nasalization of the vowel [i].

In the VDT part I the loudness contours had remedial potential for none of the elementary errors in the Articulation Substitution set. In the VDT part II, also, none of the contours of the male utterances which exemplify errors in this set were distinguishable. However, the contours of female utterances which exemplify two of the errors in this set were. These errors were EE-14 and EE-15. It should be added that the loudness contours were tested for remedial potential for only EE-13, EE-14 and EE-15 of the Articulation Substitution Set (the set comprises elementary errors EE-12 to EE-19). It was the spectral features of the target and error utterances that were expected to be predominant in this set. However due to the effects of co-articulation

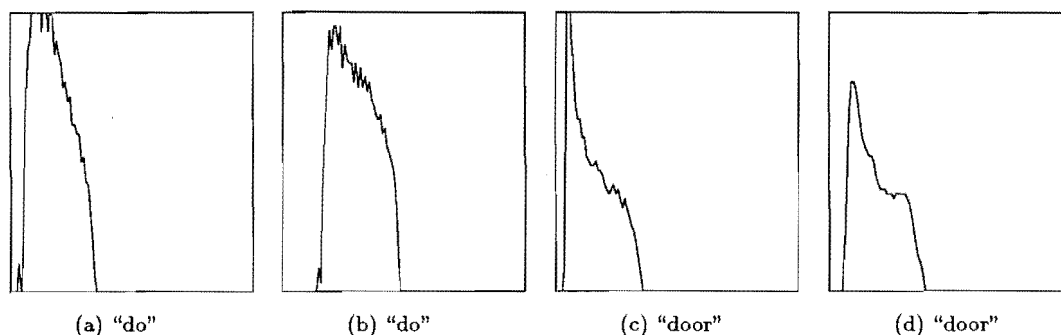


Figure 6.21. The contours of the target and error female utterances “do” and “door”, which exemplify EE-14. The contours of the target words are given in (a) and (b) and those of the error words given in (c) and (d).

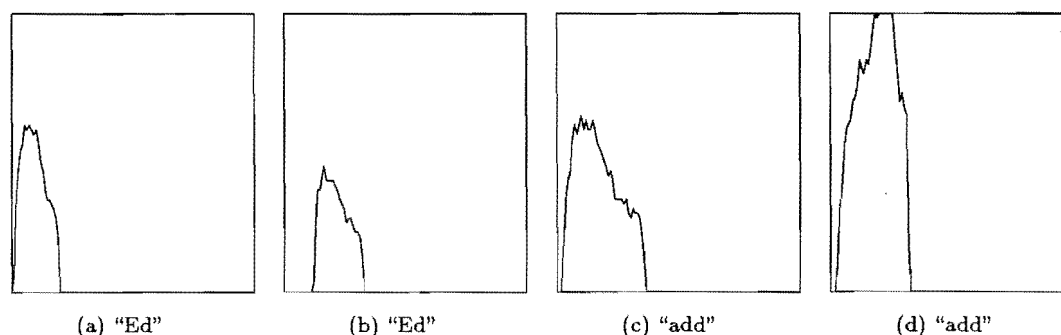


Figure 6.22. The contours of the target and error female utterances of “Ed” and “add”, which exemplify EE-15. The contours of the target words are given in (a) and (b) and those of the error words given in (c) and (d).

it was feasible that the target and error utterances of EE-13, EE-14 and EE-15 might have been distinguishable by the loudness contours, so they were tested.

Fig 6.21 shows the contours of “do” and “door”, the female utterances which exemplify EE-14. It can be seen that the contours of “door” both have an initial spike, similar to the spike in the contours of “be” in fig 6.20. The contours of “do” do not have this spike (the utterances of the error in EE-3 are the same as the target in EE-14). The contours of the target and error female utterances of “Ed” and “add”, which exemplify EE-15 are distinguishable by the width of the contours (see fig 6.22). The shape of the contours are all fairly similar. The [e] vowel in “Ed” is a short vowel and is vocalised for a shorter length of time than the [æ] vowel in “add”. It can be seen in fig 6.22 that the contours of “Ed” are not as wide as those of “add”. Thus the contours of the female utterances of “door” and “do” and of “Ed” and “add” fulfil the four requirements of remedial potential in VDT part II.

In the last two sets of speech errors, the Suprasegmental and Speech Timbre sets, the intonation and the position of the stresses in the words or phrases are important features to consider. We expected that the loudness contour display type would have remedial potential for a large number of elementary errors in these two sets. The vocal intensity of speech is related to intonation and stress. Hence any variations in these

two features should be apparent in the loudness contours.

In the VDT part I the loudness contours had remedial potential for EE-21 in the Suprasegmental set, exemplified by utterances from both the female and male speaker, (see fig 6.2 and 6.3). In the Speech Timbre set, the VDT part I revealed that the loudness contours had remedial potential for EE-23, EE-26 and EE-28, exemplified by female utterances. The contours had no remedial potential for any of the errors in this set, when exemplified by male utterances. When the loudness contours which exemplified the errors in the Suprasegmental and Speech Timbre sets were re-examined it was found that most of the contours of the target and error utterances could be distinguished if the effects of intonation and stress were noted.

In the Suprasegmental set it was possible to distinguish between the contours of the target and error utterances, spoken by either speaker, which exemplified EE-21 and EE-22. The results of VDT part I indicate it was not necessary to have knowledge of what acoustic features should be apparent in the contours of the utterances which exemplified EE-21. The loudness contours had remedial potential for EE-21 at the 90% level, when it was exemplified by female utterances and at the 80% level when the error was exemplified by male utterances. The pitch of "now?" is rising whereas the pitch of "now!" is falling. The difference in the pitch variation, results in a difference in which portion of the word "now" is intensified. The utterance "now?" is intensified at the end of the word; the utterance "now!" is intensified at the beginning. Thus the shape of the loudness contours of "now?" and "now!" are very distinctive. Hence the contours, from utterances of either speaker, fulfilled the requirements for remedial potential in VDT part II.

To distinguish between the contours of the target and error utterances which exemplified EE-22 it was necessary to know what features to look for. For this reason loudness contours did not have remedial potential for this elementary error in VDT part I (see table 6.9). The target utterance "contract" spoken as a noun was stressed on the first syllable. The error utterance of "contract" spoken as a verb was stressed on the second syllable. Figure 6.23 shows the contours of "contract" (noun) and "contract" (verb) uttered by the female speaker. By looking at the heights and widths of the contour sections, the difference between stressing the first and second syllable in the word "contract" can be seen. They could also be seen in the contours of the male utterances of "contract". Thus the distinction between the target and error utterances could be made. Hence in VDT part II the loudness contours had remedial potential for EE-22, exemplified by utterances from either speaker.

In the last set of speech errors, the Speech Timbre set, there were many different distinguishing features between the target and error utterances, aside from intonation and stress. There was the average loudness and pitch, the loudness and pitch range, the duration of the entire phrase and the vowel quality. We expect an indication of all the above mentioned features, bar those of average pitch, in the loudness contours. The most obvious feature when looking at the contours from this set is large differences between the target and error contours in duration and/or in the average loudness. The contours of the target and error female utterances which exemplified EE-23, EE-25 and EE-26 could be distinguished in this manner. The contours which exemplify EE-25 are given in fig 6.14. In the VDT part I the loudness contours had remedial potential for EE-23 and EE-26, exemplified by female utterances.

When commenting on the contours of the target male utterances of the Speech

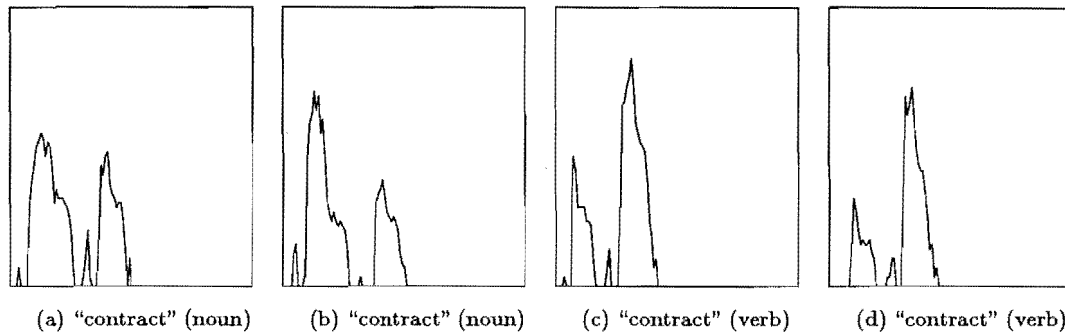


Figure 6.23. The contours of the target and error female utterances of "contract" (the noun) and "contract" (the verb), which exemplify EE-22. The contours of the target words are given in (a) and (b) and those of the error words given in (c) and (d).

Timbre set in VDT part II, an allowance must be made, since the two target contours look quite different. On examining the time domain waveforms of target utterances it was discovered that in the second the target utterance was pronounced very softly with the exception of the initial word "Grape". The loud "Grape" gave the impression the phrase was uttered louder than it actually was. Given the similarity between the female target contours (see fig 6.14) it is reasonable to speculate that if both the target phrases spoken by the male speaker had been said at the same constant vocal intensity the contours would have looked similar. Thus it will be assumed that, if both target utterances had been spoken at the same constant vocal intensity, they would have looked similar to the top contour in fig 6.24 (which is the same as the top contours in figure 6.25 too). In light of this the contours of the target and error utterances, spoken by the male speaker, which exemplify EE-24, EE-25, EE-26, EE-28 and EE-29 are distinguishable on the basis of length and/or average height. Figures 6.24 and 6.25 give the contours of the utterances which exemplify EE-26 and EE-29 respectively. The fact that the contours of the pair of error utterances in fig 6.24 look similar, as do those in fig 6.25, gives further strength to the argument that the target contours would have looked similar if they had both been said at the same constant loudness. Further support comes from the fact that in VDT part I for the elementary errors EE-23, EE-24, EE-26, EE-27, EE-28 and EE-29, exemplified by the male utterances, at least 90 % of the participants correctly identified both the error same-speech pairs in the plot-sets $X_1Y_1Y_2$ and $X_2Y_1Y_2$, see fig 6.3.

The syllables or words which were stressed in the phrase "Grape juice and water mix well" were related to the speech timbre. Stress is related to duration, pitch and loudness (Ladefoged, 1975). Thus the contour segments corresponding to the stressed syllables or words were higher and wider than when they were unstressed. The elementary errors which were distinguishable by the effects of stress on the contours were EE-24 (see fig 6.26) and EE-28 (see fig 6.27) exemplified by female utterances, and EE-23 (see fig 6.28) EE-24 and EE-28, exemplified by male utterances. In VDT part I the loudness contours only had remedial potential for EE-28 exemplified by female utterances.

Finally the vowel quality can affect the shape of the loudness contours. This is particularly obvious for nasal and breathy speech. The contours of the target and error utterances, spoken by either speaker, which exemplified EE-27 and EE-28 were quite

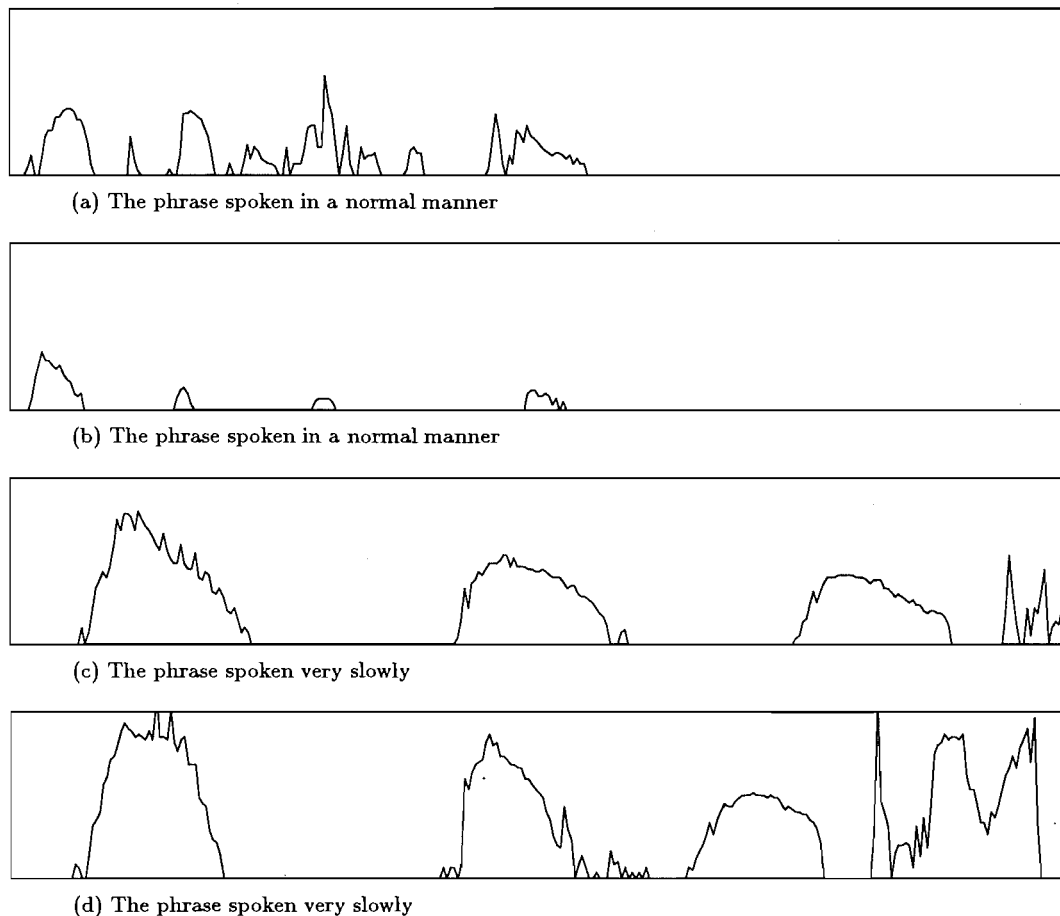


Figure 6.24. The contours of the target and error male utterances, which exemplify EE-26. The contours of the target phrases “Grape juice and water mix well” spoken in a normal voice are given in the top two displays. The contours of the error, the above mentioned phrase spoken very slowly, are given on the bottom two displays.

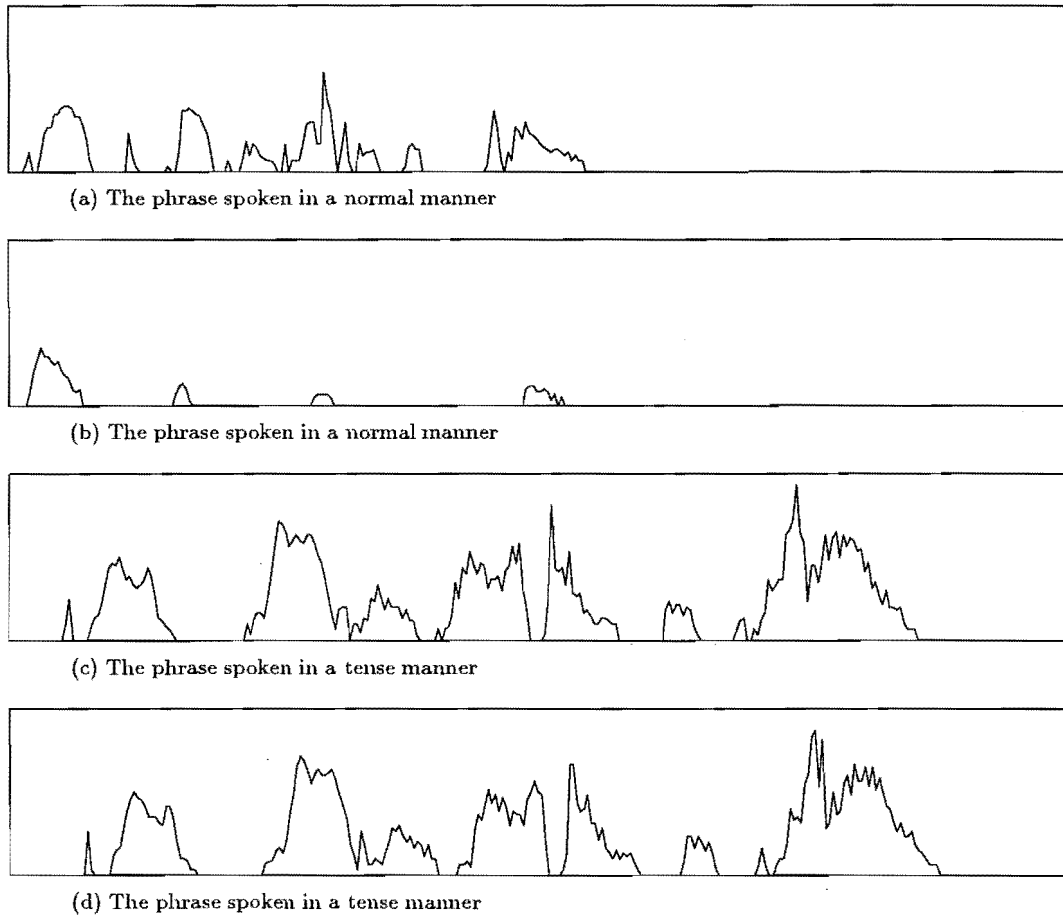


Figure 6.25. The contours of the target and error utterances, which exemplify EE-29. The contours of the target phrases "Grape juice and water mix well" spoken in a normal voice are given in the top two displays. The contours of the error, the above mentioned phrase spoken in a tense manner, are given on the bottom two displays.

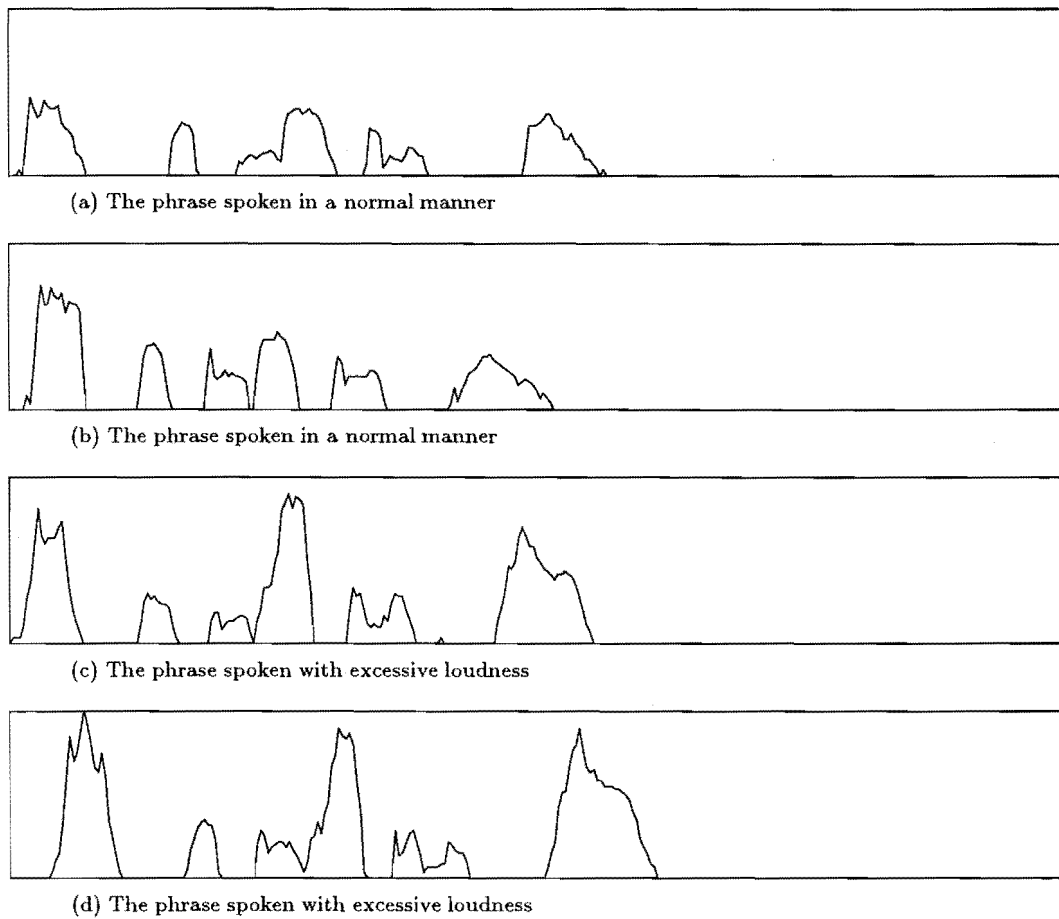


Figure 6.26. The contours of the target and error female utterances, which exemplify EE-24. The contours of the target phrases "Grape juice and water mix well" spoken in a normal voice are given in the top two displays. The contours of the error, the above mentioned phrase spoken with excessive loudness, are given on the bottom two displays.

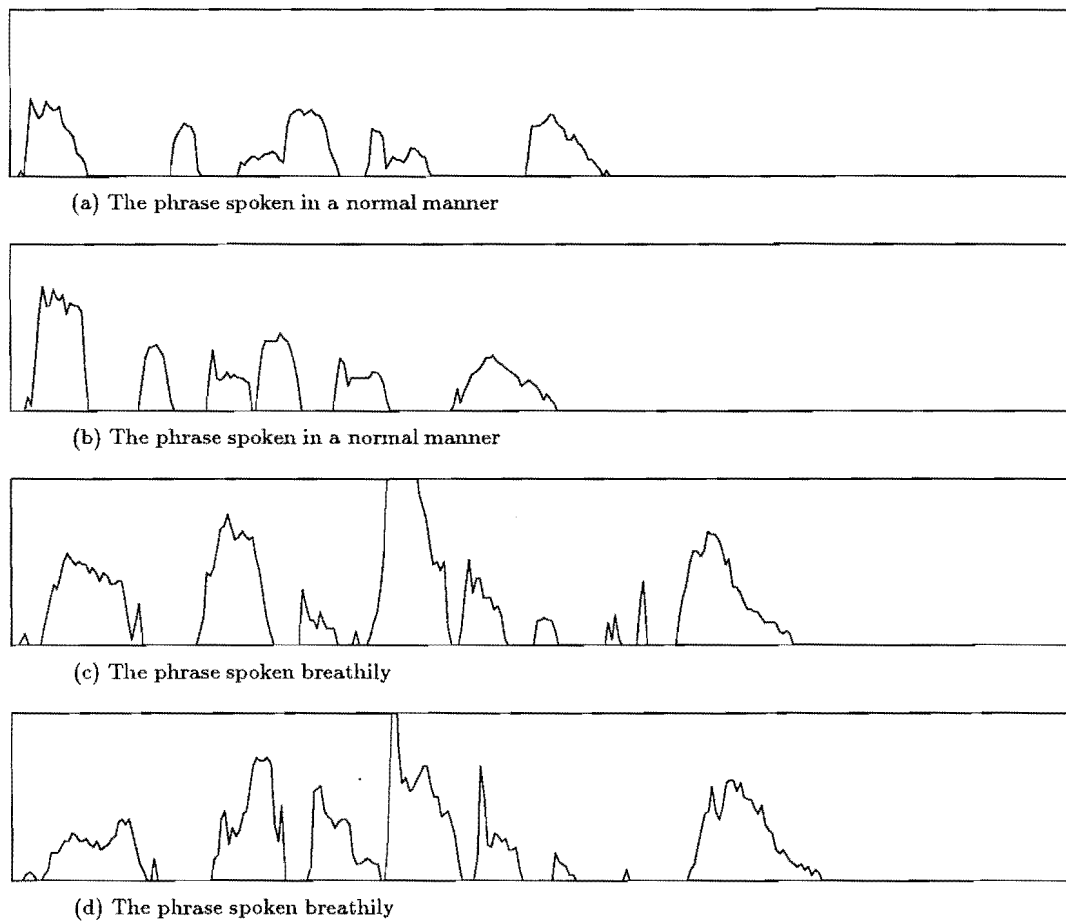


Figure 6.27. The contours of the target and error female utterances, which exemplify EE-28. The contours of the target phrases "Grape juice and water mix well" spoken in a normal voice are given in the top two displays. The contours of the error, the above mentioned phrase spoken breathily, are given on the bottom two displays.

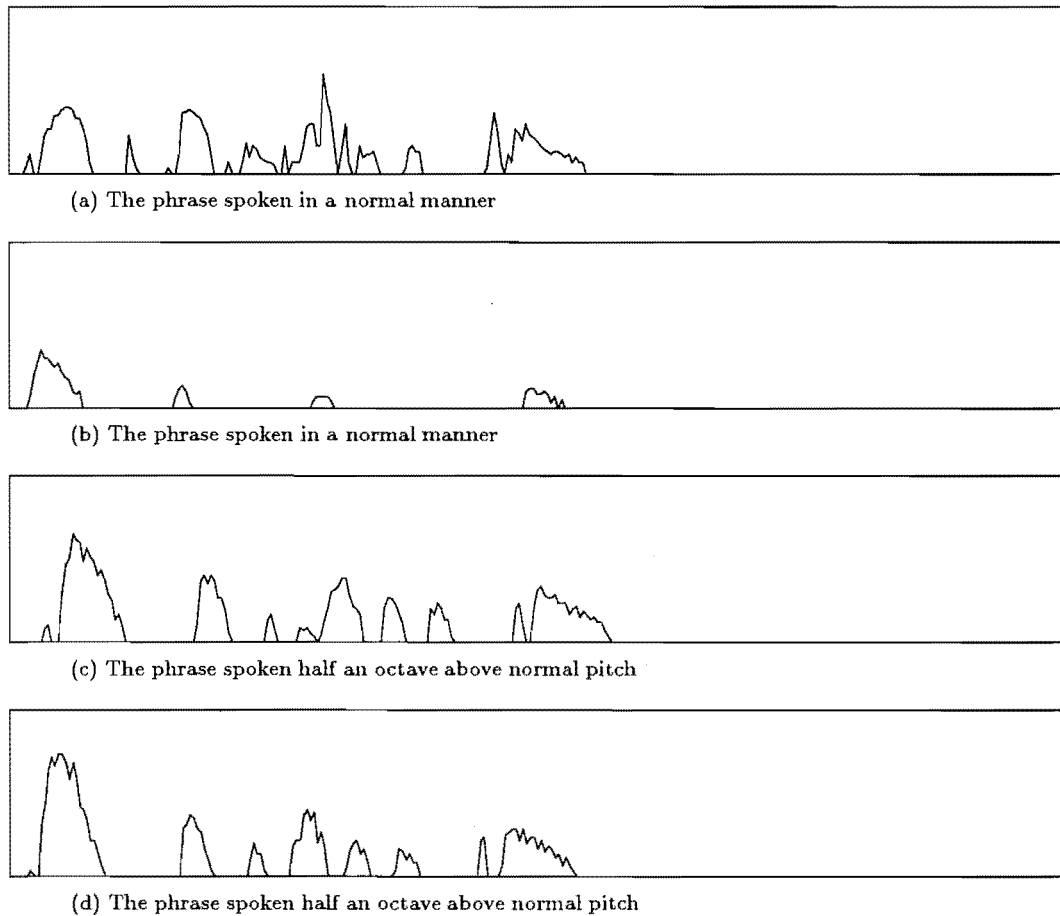


Figure 6.28. The contours of the target and error male utterances, which exemplify EE-23. The contours of the target phrases “Grape juice and water mix well” spoken in a normal voice are given in the top two displays. The contours of the error, the above mentioned phrase spoken half an octave above normal pitch, are given on the bottom two displays.

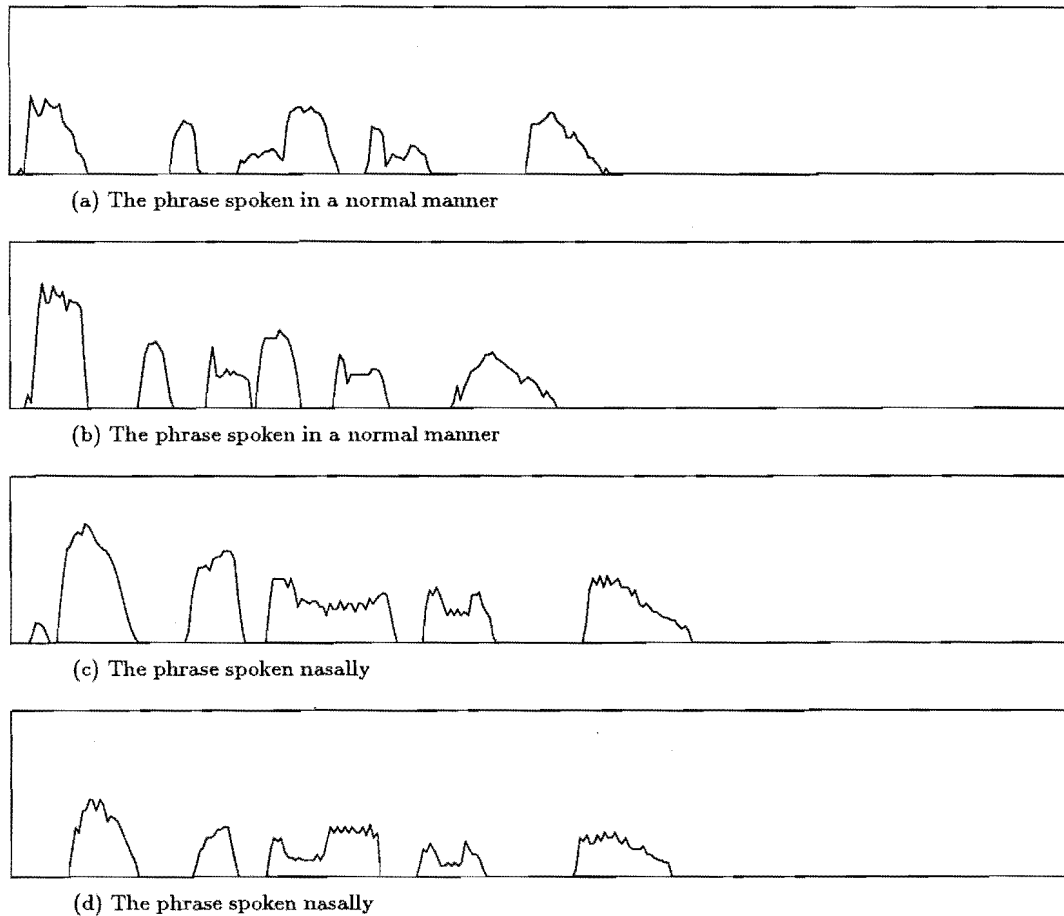


Figure 6.29. The contours of the target and error female utterances, which exemplify EE-27. The contours of the target phrases "Grape juice and water mix well" spoken in a normal voice are given in the top two displays. The contours of the error, the above mentioned phrase spoken nasally, are given on the bottom two displays.

distinctive by shape, for example see fig 6.29 and 6.27. The contours in fig 6.29 exhibit wiggly lines in some of the high portion of the contours. The wiggles are similar to those seen in the contours of "me" in fig 6.20. This suggests the wiggles are evidence of nasalization.

The VDT part II revealed that the loudness contours were a very useful speech tool (see table 6.9). This suggests that we can therefore accept the evidence provided by the therapists that the Loudness Monitor is an effective module (see sec 5.2.1). It was interesting to see that every elementary error the loudness contours had remedial potential for in VDT part I they also had remedial potential for in VDT part II. In VDT part I the loudness contours only had remedial potential for 7 (out of a possible 23) elementary errors, exemplified by female utterances, but the contours had remedial potential for 15 (out of a possible 23) elementary errors in VDT part II. For the elementary errors, exemplified by male utterances, the loudness contours had remedial potential for a paltry 3 (out of a possible 23) elementary errors in VDT part I and 15 (out of a possible 23) elementary errors in VDT part II. Two conclusions can be drawn from these results. Firstly for the elementary errors for which the loudness contours had remedial potential in VDT part I the most obvious difference between the target and error displays *was* related to the intended acoustic difference between the target

| TEST | UTTERANCE | ART. INT. | | | UN./V. | | N. | ART. SUB. | | | | SUPR. | | | SP. TIMB. | | | | |
|--------|-----------|-----------|---|---|--------|---|----|-----------|----|----|----|-------|----|----|-----------|----|----|----|----|
| | | 2 | 3 | 4 | 6 | 8 | 11 | 14 | 15 | 17 | 19 | 20 | 21 | 22 | 23 | 26 | 27 | 28 | 29 |
| VDT II | Female | | | ✓ | | | | | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| VDT II | Male | | | ✓ | ✓ | | ✓ | | ✓ | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| VDT I | Female | | | ✓ | | | ✓ | | | | | | ✓ | | ✓ | | | | |
| VDT I | Male | | | ✓ | | | | | ✓ | | | | ✓ | | | | | | |

Table 6.10. The Elementary Errors for which the pitch contours had remedial potential in VDT part I and II are indicated by a tick.

and error utterances. Therefore for this display type the implicit assumption of VDT part I held. The second conclusion is that knowledge of what display features one is supposed to look for is important, as this increased the number of elementary errors for which the loudness contours had remedial potential.

It was interesting to see that in the VDT part II there were 11 elementary errors exemplified by both female and male utterances for which the loudness contours had remedial potential, while in the VDT part I there were only two (see table 6.9). Thus in the VDT part II for most of the elementary errors for which the loudness contours have remedial potential the same acoustic features can be accounted for in the contours of both the female and male utterances.

Two features of the contours which the participants in VDT part I consistently overlooked were the width and height of the contours. These are two important features of the loudness contours, as was revealed in VDT part II. Clearly these features need to be made more obvious. One possible solution would be to shade in the region beneath the contours. A vertical scale and an adjustable horizontal line (as in the Loudness Module on the CASTT) would also increase awareness of height.

6.5.1.2 The Displays Of Pitch Contours

Table 6.10 gives the elementary errors for which the pitch contours had remedial potential in VDT part II. It is divided into the six speech error sets of the Elementary Error list, given in tables 6.1 and 6.2. For comparison the elementary errors for which the contours had remedial potential for in VDT part I have also been included. Each of the elementary errors for which the pitch contours have remedial potential are indicated by a tick. It can be seen immediately from table 6.10 that the pitch contours have remedial potential for many more elementary errors in VDT part II than in VDT part I.

In the first four speech error sets in the elementary error list (the Articulation Intensity Set, The Unvoiced/Voiced Set, the Nasal set and the Articulation Substitution set) there were only two types of elementary errors for which the pitch contours were expected to have any possible remedial potential. These were the elementary errors in which the target and error utterances differed in the length of voicing or in the number of syllables. The pitch algorithm (see sec 4.4) only estimates the pitch on voiced segments of speech (there is no period component in unvoiced speech from which the pitch could be estimated). Thus words which differ noticeably in the length of voicing can be distinguished by the length of the contours. Words which differ in the number of syllables differ in the portion of the word which is stressed. More stressed sounds are spoken louder and longer than less stressed sounds. The latter feature is made

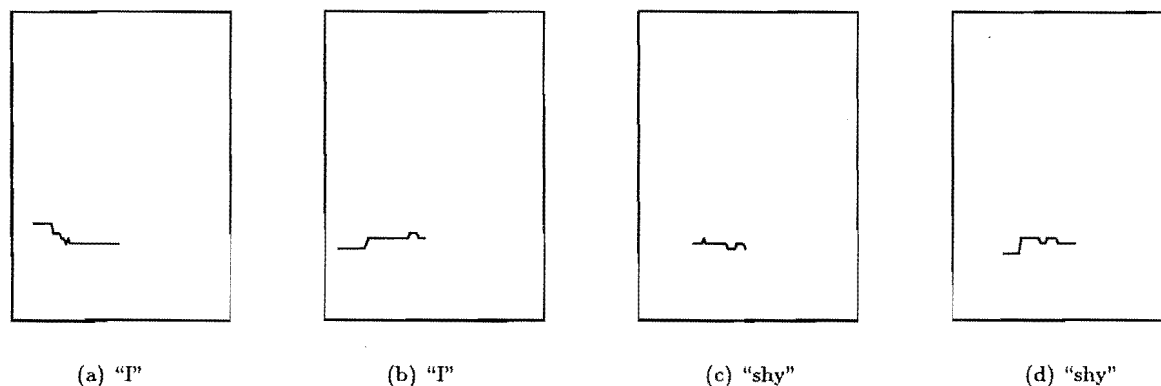


Figure 6.30. The pitch contours of target and error male utterances of “I” and “shy”, which exemplify EE-6. The contours of the target utterances are given in (a) and (b), the ones of the error utterances are given in (c) and (d).

apparent by the length of the pitch contours.

Most variations in the gradient of the pitch contours is due to intonation. Intonation is used to convey meaning and can change depending on context. In the elementary errors of the Suprasegmental and Speech Timbre sets the intonation of the words and phrases is an important feature. However it plays no part in distinguishing between the target and error utterances which exemplify EE-1 through to EE-19. A quick glance at the contours of these utterances revealed that the gradients of these contours varied from utterance to utterance. The intonation pattern imposed on these words is most likely to be due to the recording style. The words were recorded in groups of three. Hence the gradient of the pitch contours in EE-1 through to EE-19 was not considered to be an important feature.

The pitch target and error contours of EE-4 and EE-15 exemplified by utterances from either speaker and of EE-6, EE-11 and EE-17 exemplified by male utterances could all be distinguished by the length of the contours. In each of these cases the difference in the contour length directly corresponds to the voicing duration. The elementary errors EE-6, EE-15 and EE-17 are all exemplified by one-syllable target and error utterances. Thus the differences in contour length are due to the differences in the voicing duration of different sounds and to co-articulation effects.

Figure 6.30 gives the contours of the male utterances of “I” and “Shy”, the combination which exemplifies EE-6. A voiced sound (such as [ai]) when preceded by an unvoiced sound (such as [ʃ]) is often shorter than when it is not. This can be seen in fig 6.30, as the contours of the diphthong [ai] in “I” are longer than those of “shy” (since [ʃ] is unvoiced, no pitch value is calculated for it and hence there is no pitch contour for it). The target and error utterances which exemplify EE-15 are “Ed” and “add”. The vowel [e] in “Ed” is voiced for a shorter length of time than the vowel [æ] in add. The difference in the length can be seen in the length of the contours for “Ed” and “add”, spoken by either speaker. Figure 6.31 is the contours of the female utterances of “Ed” and “add”. Although not shown, the difference in voicing length between [ju] in “use” and [u] in “ooze” could also be seen by the length of the contours of the male utterances (“use” and “ooze” are the target/error combination which exemplify EE-17). The contours of “use” are longer than those for “ooze”.

Thus in the VDT part II the pitch contours had remedial potential for EE-15

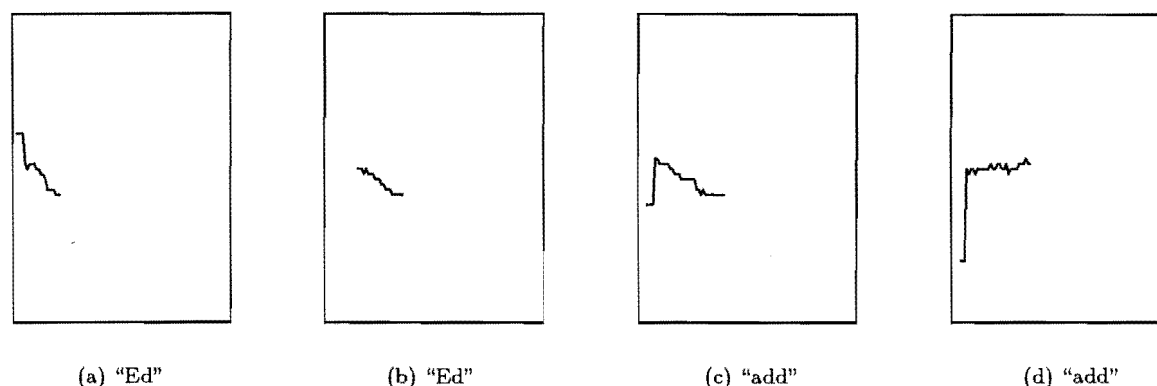


Figure 6.31. The pitch contours of target and error female utterances of "Ed" and "add", which exemplify EE-15. The contours of the target utterances are given in (a) and (b), the ones of the error utterances are given in (c) and (d).

exemplified by utterances of either speaker and for EE-6 and EE-17 exemplified by male utterances (see table 6.10). All of these target and error contours could be distinguished on the basis of length only. The gradient of the contours was of no consequence and was merely a distracting feature (see fig 6.30 and 6.31). There was one instance, the contours of the male utterances which exemplify EE-15, where the contours of the target and error utterances could be distinguished on the basis of length and by the gradients. However this was just mere coincidence. These contours were the only ones, in the above mentioned group, in which in VDT part I the pitch contours had remedial potential (see fig 6.4 and fig 6.5). Clearly with no knowledge of what features to look for the participants regarded the gradient of the contours as an important feature.

In EE-1 through to EE-19 there were only two instances in which the target and error utterances differed in the number of syllables. These were the utterances which exemplified EE-4 and EE-11. In most cases the difference between words which differed in syllables could be seen by the length of the pitch contours. The difference in length between the vowel [ɒ] in "poppa" and "pop" could be seen by the length of the contours of both the female and male utterances ("poppa" and "pop" exemplify EE-4). The sound [pɒ] is more stressed in "pop" than in "poppa" (see sec 6.5.1.1), so the contours are longer. This can be seen in fig 6.32, for the contours of the female utterances of "poppa" and "pop". Not only are the differences in length apparent but it can be seen that the contour segment of [pɒ] in "pop" has a positive slope, whereas in "poppa" it is flat. The same features were observed in the contours of the male utterances. Thus in the VDT part II the pitch contours had remedial potential for EE-4 exemplified by either male or female utterances. The distinctions between the target and error contours were so obvious that in VDT part I the pitch contours had remedial potential at the 80 % level for both the female and male utterances which exemplified EE-4.

The target and error contours of the male utterances of "me" and "mbee" (these exemplify EE-11) were also distinguishable by length. The contours of "mbee" were longer than those for "me". However the contours of female utterances were not distinguishable. Ironically the pitch contours had remedial potential in VDT part I for EE-11 exemplified by female utterances but not male utterances.

Figure 6.33 shows the contours of the female utterances of "me" and "mbee". On looking at those contours it is most likely that the visual feature the participants in VDT

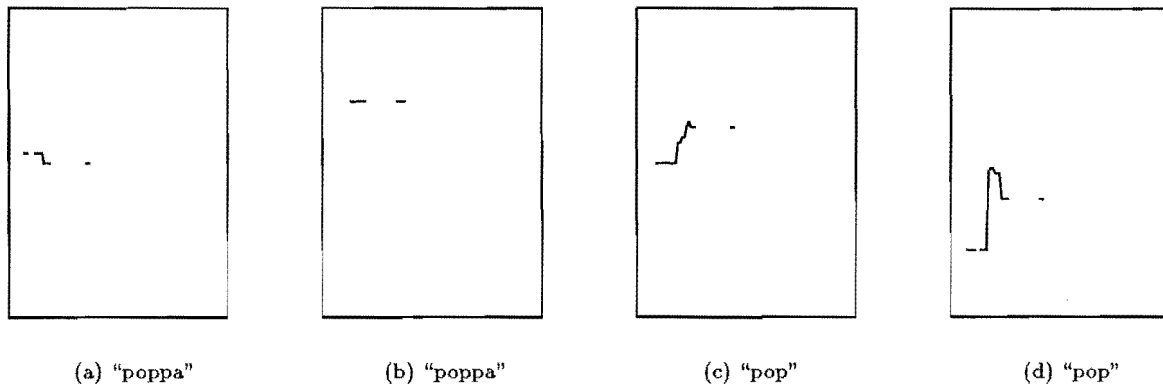


Figure 6.32. The pitch contours of target and error female utterances of "pop" and "poppa", which exemplify EE-4. The contours of the target utterances are given in (a) and (b), the ones of the error utterances are given in (c) and (d).

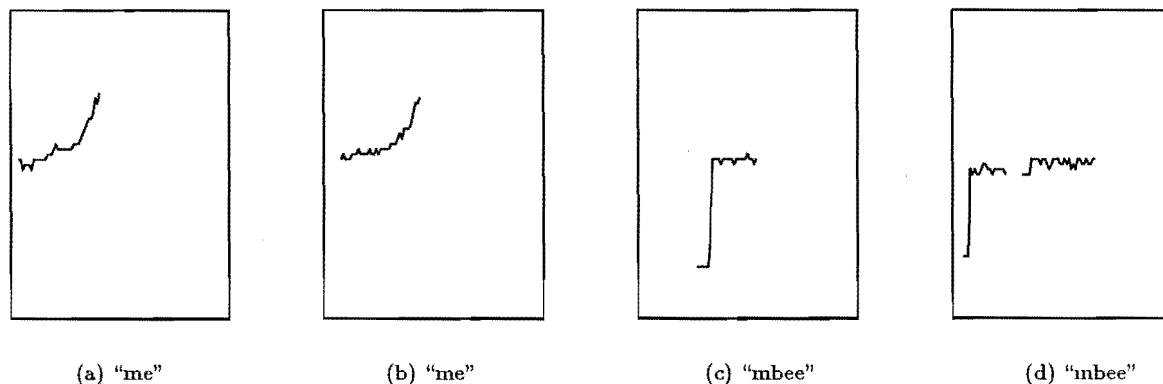


Figure 6.33. The pitch contours of target and error female utterances of "me" and "mbee", which exemplify EE-11. The contours of the target utterances are given in (a) and (b), the ones of the error utterances are given in (c) and (d).

part I used to separate the target and error utterances was the fact that the contours of "mbee" had a step in them whereas the contours of "me" had a positive slope. It is not likely that these features are related to the acoustic differences between "me" and "mbee". The step in the "mbee" contours is mostly likely due to the inadequacies of the pitch tracking algorithm. The algorithm has great difficulty tracking the pitch initially in an utterance. Often the initial calculated pitch value is a frequency well below the actual pitch value. Many of the other pitch contours from the target and error utterances exhibited this step feature, for example fig 6.31. This distracting feature in all the pitch contours should always be ignored. The contours of "me" and "mbee" in fig 6.33 cannot be distinguished by length even if the step and the positive slope in the contours are disregarded. Thus the pitch contours do not have remedial potential for EE-11 exemplified by female utterances.

The pitch contours had remedial potential in VDT part II for most of the elementary errors in the Suprasegmental and Speech Timbre error sets. This was expected. In these two sets the intonation and the position of stresses in words or phrases are important features to be considered. Pitch is related to intonation and stress. The pitch contours were not tested for remedial potential of EE-24 and EE-25. Both these speech errors were to do with average loudness; the pitch contours were not expected

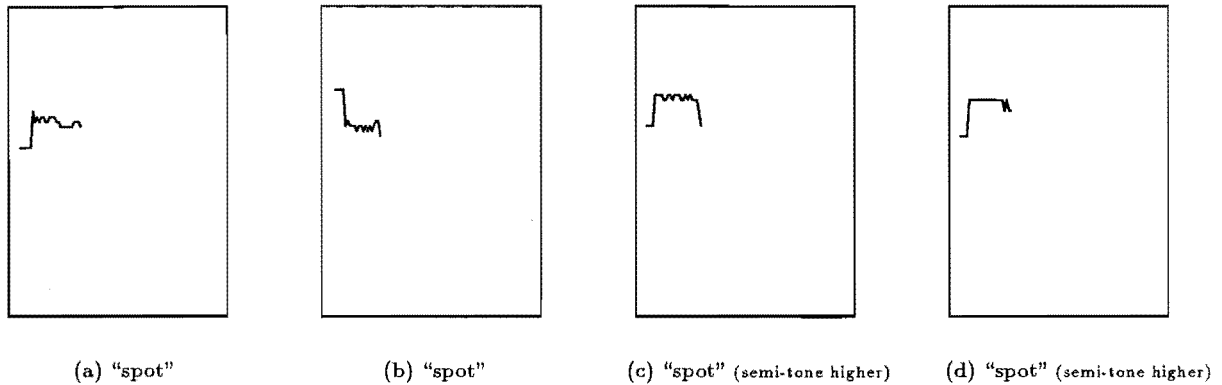


Figure 6.34. The pitch contours of target and error female utterances of "spot" and "spot" spoken a semi-tone apart, which exemplify EE-20. The contours of the target utterances are given in (a) and (b), the ones of the error utterances are given in (c) and (d).

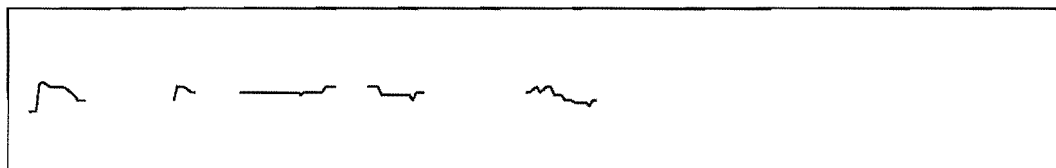
to show differences in this speech feature.

In the Suprasegmental error set the pitch contours had remedial potential for EE-20, exemplified by female utterances and EE-21 and EE-22, exemplified by utterances of either speaker. The utterances of "spot" and "spot" were spoken a semi-tone apart (these exemplify EE-20). The height of the pitch contours is proportional to the pitch so changes in pitch should result in changes in contour height. It is possible to distinguish between contours of the female utterances of "spot" and "spot" spoken a semi-tone apart on this basis. This was not obvious at first glance however because of the step effect, due to the inadequacies of the pitch tracking algorithm, on some of the contours (see fig 6.34).

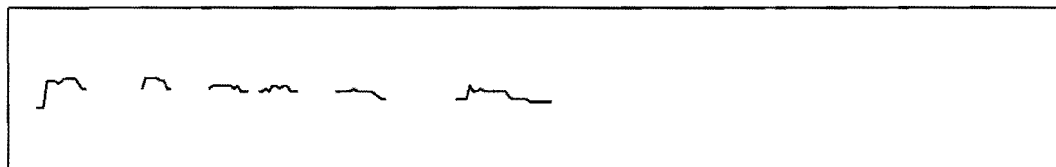
The difference between "now?" and "now!" (the target and error utterances which exemplify EE-21) is the pitch variation. The utterance "now?" finishes on a rising tone whereas "now!" finishes on a falling tone. This difference is apparent in the gradient of the contours. It was possible to distinguish between the target and error utterances of either speaker by considering the gradient. The pitch contours not only had remedial potential for EE-21, exemplified by either female or male utterances in VDT part II but the contours also had remedial potential in VDT part I. In the VDT part I the pitch contours had remedial potential at the 90% level for the female utterances and at the 80% level for the male utterances (see fig 6.4 and 6.5). Thus no knowledge was needed to distinguish between the target and error contours from the female or male utterances.

It was possible to distinguish the contours of "contract" (noun) and "contract" (verb) spoken by either speaker. It could be seen, by the length of the contour segments that the syllable "con" is stressed in the noun and the syllable "tract" is stressed in the "verb". Hence in VDT part II the pitch contours had remedial potential for EE-22 exemplified by either female or male utterances.

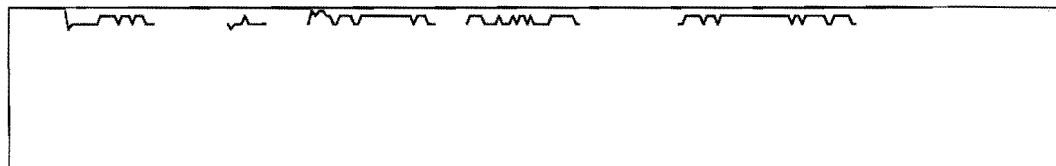
For the Speech Timbre set in VDT part II the pitch contours had remedial potential for EE-23, EE-26, EE-27 and EE-28 exemplified by utterances from either speaker and for EE-29 exemplified by male utterances. There were two main features which enabled distinction between these target and error contours. These were the average height of the contours and the length of the contours (the length of either a segment or of the entire contour).



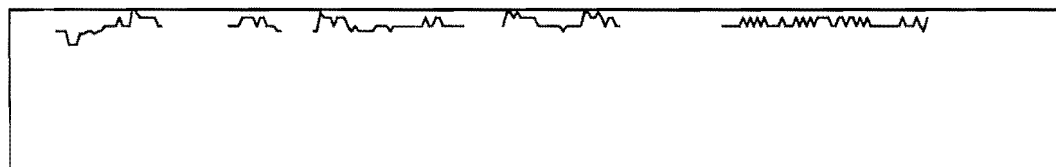
(a) The phrase spoken in a normal manner



(b) The phrase spoken in a normal manner



(c) The phrase spoken half an octave higher than normal pitch



(d) The phrase spoken half an octave higher than normal pitch

Figure 6.35. The pitch contours of the target and error female utterances, which exemplify EE-23. The contours of the target phrases "Grape juice and water mix well" spoken in a normal voice are given in the top two displays. The contours of the error, the above mentioned phrase spoken an octave higher, are given on the bottom two displays.

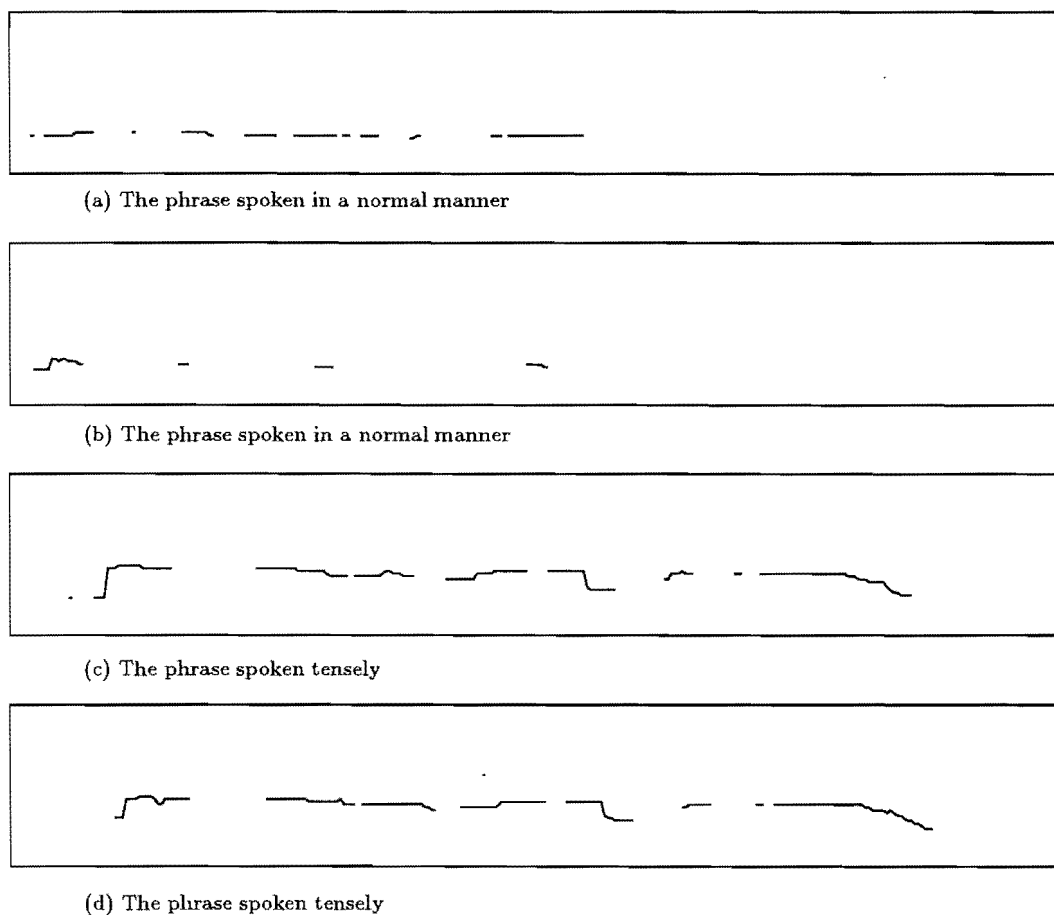


Figure 6.36. The contours of the target and error male utterances, which exemplify EE-29. The contours of the target phrases "Grape juice and water mix well" spoken in a normal voice are given in the top two displays. The contours of the error, the above mentioned phrase spoken tensely, are given on the bottom two displays.

It was possible to distinguish between the target and error contours of EE-23 (see fig 6.35), EE-26 and EE-28 exemplified by female utterances by the length of the contours and the differences in which words in the phrase were stressed. This could also be done for the contours of the male target and male error utterances which exemplify EE-26, EE-27, EE-28 and EE-29 (see fig 6.36). The length of the contours was affected by the speaking rate and by the number of words in the phrase which were stressed. Both of these two features are affected by voice timbre. Figure 6.35 shows the contours of the female utterances of the phrase spoken at normal pitch and half an octave higher. It can be seen that the contours of the phrase spoken half an octave higher than normal pitch are longer than the contours for the phrase spoken at normal pitch. It can also be seen that each word in the error utterances are stressed unlike those in the target utterances.

The ability to distinguish between the target and error contours on the basis of length can also be seen in fig 6.36, the phrase spoken normally and tensely by the male speaker. In order to do this, however, it is necessary to make one allowance. It can be seen in fig 6.36 that the contour of the second target utterance contains very few contour segments. This was the utterance discussed in sec 6.5.1.1, which was spoken very softly with the exception of the initial word "Grape". It can be seen in figure 6.24 that portions of this utterance were so quietly spoken that no loudness contours were calculated for it, most of the speech being unvoiced. Therefore there were very few segments of speech for which the pitch could be calculated. However it can be seen that the average heights of the two contours are similar which means the average pitch of the utterances were similar. By investigating the original time domain waveforms it was seen that the duration of the two phrases was also similar.

Changing the speech timbre can also change the average pitch of the voice. Changes in average pitch would be seen as changes in the average height of the pitch contours. By comparing the average heights of the target and error contours we found the target and error contours of EE-23 and EE-27 exemplified by female utterances to be distinguishable by height, and also the contours of EE-23 and EE-29 exemplified by male utterances.

Whilst most of the target and error contours of the Speech Timbre set were clearly distinguishable by height or length the pitch contours only had remedial potential for EE-23 exemplified by female utterances in VDT part I. The fact that most of the participants in VDT part I did not correctly identify all the same speech pairs from the displays is mostly probably due to the inadequacies of the pitch tracking algorithm. The steplike effects in the contours due to the incorrect initial pitch value.

The pitch range is also affected by speech timbre. However it was not possible to use pitch range as a feature to distinguish between the target and error utterances of the Speech Timbre set. This was due to the linear frequency scale of the displays in which the contours were plotted. The pitch scale is nonlinear. The frequency range of an average male voice whose pitch values vary over an octave will be much much less than the frequency range of an average female voice whose pitch values vary over an octave. The contours did not show the pitch variations of the female speaker very well let alone those of the male speaker with the lower pitch. It was only when the pitch variations were exaggerated as in "now?" and "now!" that they were clearly shown on the contours.

The results of the VDT part II showed that once again the implicit assumption of

the VDT part I held for the pitch contours (the implicit assumption being that the most obvious difference between the target and error displays could directly be attributed to the acoustic difference between the target and error utterances). In all but one instance the elementary errors for which the pitch contours had remedial potential in VDT part I, the contours also had remedial potential in VDT part II. The number of elementary errors for which the pitch contours had remedial potential increased in VDT part II from VDT part I. It can be seen in table 6.10 that in VDT part II the pitch contours have remedial potential for 9 (out of a possible 18) elementary errors exemplified by female utterances and 11 (out of a possible 18) elementary errors exemplified by male utterances. This is clearly more elementary errors than in VDT part I where the pitch contours had remedial potential for 4 (out of 18) elementary errors exemplified by female utterances and 3 (out of 18) elementary errors exemplified by male utterances. In addition, for most of the elementary errors for which the pitch contours had remedial potential, the elementary errors were exemplified by both female and male utterances. Thus in VDT part II there were eight elementary errors exemplified by utterances from both speakers, for which the contours had remedial potential. The results of the VDT suggest the evidence provided by the therapists that the pitch contours are an effective speech tool (see sec 5.2.1), can therefore be accepted.

Contrasting the VDT part I and VDT part II results highlighted a few areas where the pitch contours display could be improved in several ways. The pitch tracking algorithm must be improved in order that the initial calculated values of the pitch are correct. If they were this would get rid of the step effect, which was the most distracting feature in the displays in the VDT. The pitch range in the current pitch contours of an utterance is not a very noticeable feature, except when the pitch variation in an utterance has been exaggerated. Pitch range, or lack of it, is an important feature of speech timbre. More emphasis would be placed on the pitch range if in the pitch displays there were separate frequency scales for high pitched voices (usually women's and children's) and for low pitched voices (usually men's).

The participants in VDT part I consistently ignored height as an important feature, as they did with the loudness contours. The solutions suggested for the loudness contours - colouring the region beneath the contours, an adjustable vertical line and a frequency scale - are possible solutions to increase the user's awareness of the heights of the pitch contours too. Once the pitch algorithm is improved and the step effect does not occur any more, the differences in the height of the pitch contours may automatically become more obvious anyway.

6.5.1.3 The Spectral Content Display

Table 6.11 gives the number of elementary errors for which the spectral content plots had remedial potential in VDT part II. For comparison purposes the elementary errors for which the plots had remedial potential in VDT part I have also been included in the table. It can be seen that, as with the pitch and loudness contours the spectral plots had remedial potential for far more elementary errors in VDT part II than in VDT part I.

The spectral plots were tested for remedial potential for all the speech errors in the Articulation Intensity, Nasality and Articulation Substitution sets and for most of the speech errors in the Unvoiced and Suprasegmental sets. The target and error

| TEST | UTTERANCE | ART. INT. | | | | | UN./V. | | | NAS. | | ART. SUB. | | | | | | | | | | SUPR. | |
|--------|-----------|-----------|---|---|---|---|--------|---|---|------|----|-----------|----|----|----|----|----|----|----|----|----|-------|--|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 21 | 22 | | |
| VDT II | Female | ✓ | | ✓ | ✓ | | ✓ | ✓ | | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | | |
| VDT II | Male | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | | | ✓ | ✓ | ✓ | | |
| VDT I | Female | ✓ | | | ✓ | | | ✓ | | | | ✓ | ✓ | | ✓ | ✓ | | ✓ | | | | | |
| VDT I | Male | ✓ | | | | | | | | | | | ✓ | | | | | | | | | | |

Table 6.11. The Elementary Errors for which the spectral plots had remedial potential in VDT part I and II are indicated by a tick.

utterances in the Articulation Intensity set through to the Articulation Substitution set (EE-1 to EE-19) all differed in one phoneme. The elementary errors EE-3, EE-5 to EE-12, EE-15, EE-18 and EE-19 all differed in the initial phoneme. The errors EE-2, EE-16 and EE-17 differed in the medial phonemes and EE-1, EE-4, EE-13 and EE-14 differed in the final phoneme. These facts indicate where in the target and error plots we should be looking for differences.

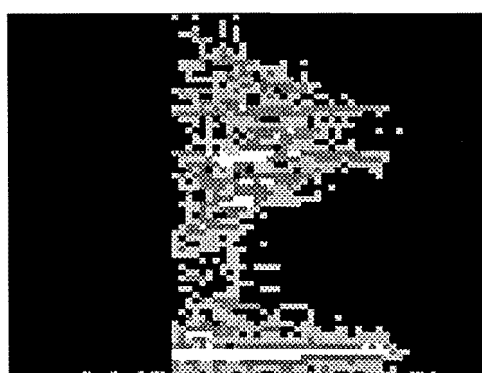
The spectral plots look more like multi-coloured shapes on a black background than traditional spectrograms. Since the CASTT had a CGA graphics card, only four colours were available for the spectrogram plots (see sec 4.2). Due to this the energy level was quantised into four regions. The four colours of the spectral plots in the VDT were black, red, blue and white. The frequency range of the spectrogram for the VDT was from 0 - 4.5 kHz (note that this was larger than the bandwidth of 0 - 3.0 kHz of the spectrograms in the Spectrogram Module on the CASTT because of the different filter characteristics of the low-pass filters on the TMSboard and the SX8 board, see sec 6.3.3).

Despite the crudeness of the spectrogram we were able to get quite distinct plots for many of the target and error utterances. Figure 6.37 shows the spectral plots of the female utterances “see” and “she” (the target and error utterances which exemplify EE-12). All the spectral plots shown in this section are in a grey scale rather than coloured as in the original plots used in the VDT. The regions of the original spectral plots that were coloured black are still black; similarly the regions that were coloured white are still white. However the regions that were coloured blue in the original plots are now light grey and the regions that were coloured red are now dark grey.

The colour assigned to the regions where the frequencies are of low energy was black. Since this is the same as the background colour the spectral plots look like shapes. It can be seen in fig 6.37 that the shapes of target plots are quite distinct from the shapes of the error plots.

It is very important to note that many of the plots of target and error utterances are quite distinctive on the basis of *shape*. The distribution of the *colours* within the spectral plots is also a feature to be aware of, especially the white regions. These are the regions where the energy of the frequency components is the highest. Figures 6.37, 6.38 and 6.39 are three examples in which the target and error plots can be distinguished by shape. It can be seen in fig 6.37 that in the lower right half of all the spectral plots there is a sizeable white strip. The second half of both the utterances of “see” and “she” is [i]. It can be seen that the second half of all four plots, corresponding to [i], in fig 6.37 are very similar.

The difference between “see” and “she” is in the initial phonemes. The most no-



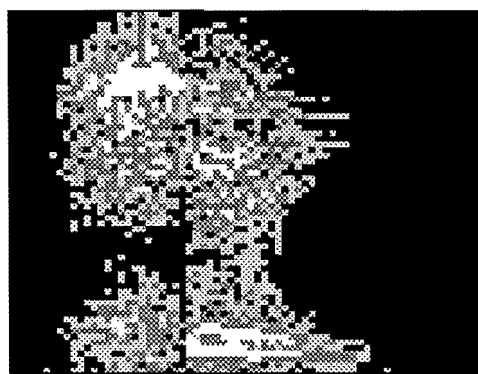
(a) "see"



(b) "see"



(c) "she"



(d) "she"

Figure 6.37. The spectral plots of target and error female utterances of "see" and "she", which exemplify EE-12. The spectral plots of the target utterances are given in (a) and (b), the ones of the error utterances are given in (c) and (d).

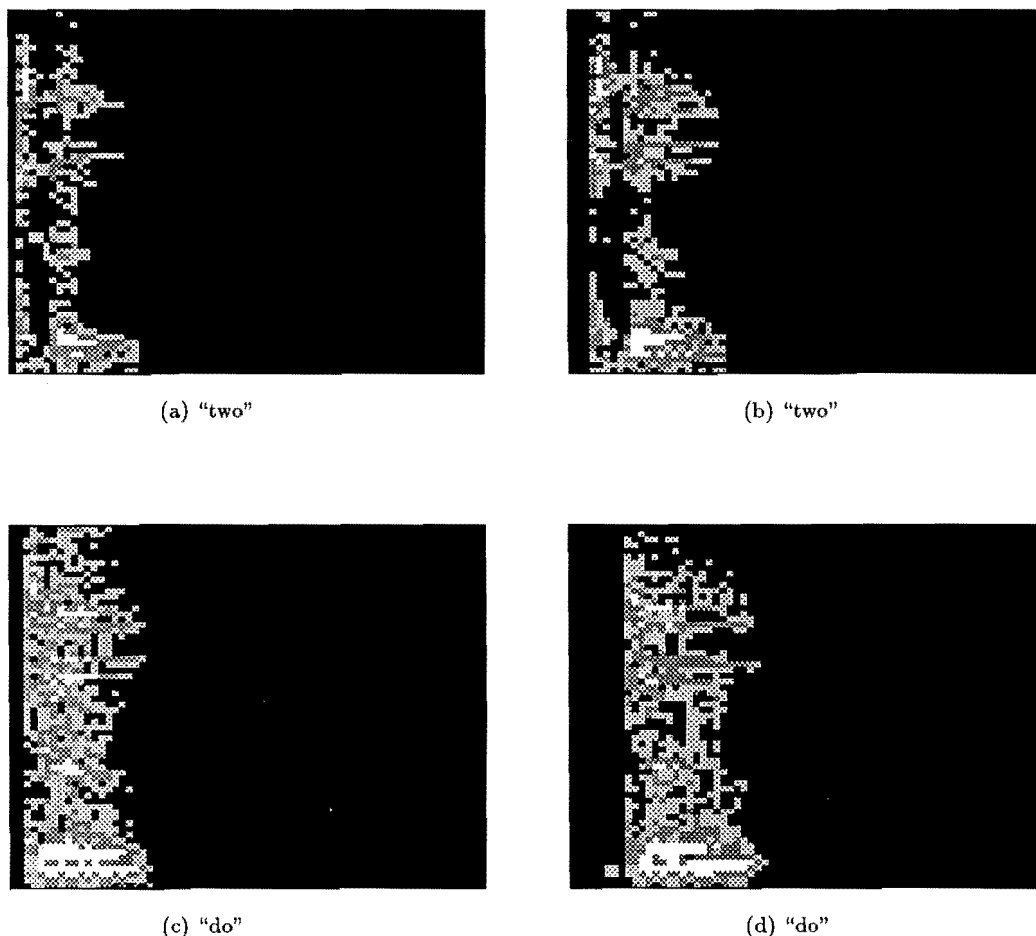


Figure 6.38. The spectral plots of target and error female utterances of "two" and "do", which exemplify EE-3. The spectral plots of the target utterances are given in (a) and (b), the ones of the error utterances are given on the bottom two.

ticeable difference between the target and error plots in fig 6.37 is the initial region. The [f] sound has a high energy concentration in the 2-4 kHz region (see sec 7.4.1). This can be seen in fig 6.37 by the white region in the top left half of the plots. The sound [s] has a high energy concentration in the 5 - 8 kHz region. This is well above the frequency range of the spectrogram (0 - 4.5 kHz). The energy below 5 kHz is very low, as can be seen in fig 6.37 since the left half of the plots for [s] is virtually black. The reason for the presence of a few light grey squares in the top left region of the second utterance of "see" will be discussed soon.

Figure 6.38 is another example of how the plots can be distinguished on the basis of shape. Figure 6.38 gives the plots of the female utterances of "two" and "do" (the target and error utterances which exemplify EE-3). The difference between "two" and "do" is in the initial phonemes, [t] being an unvoiced alveolar plosive and [d] being its voiced pair. The voice onset time will be greater for "two" than "do" due to the initial unvoiced plosive. This can be seen in fig 6.38. There is a region of low energy, indicated by the vertical black strip, after the initial [t] in the contours of "two". This region cannot be seen in the contours of "do". The right half of the plots, of "two" and "do",

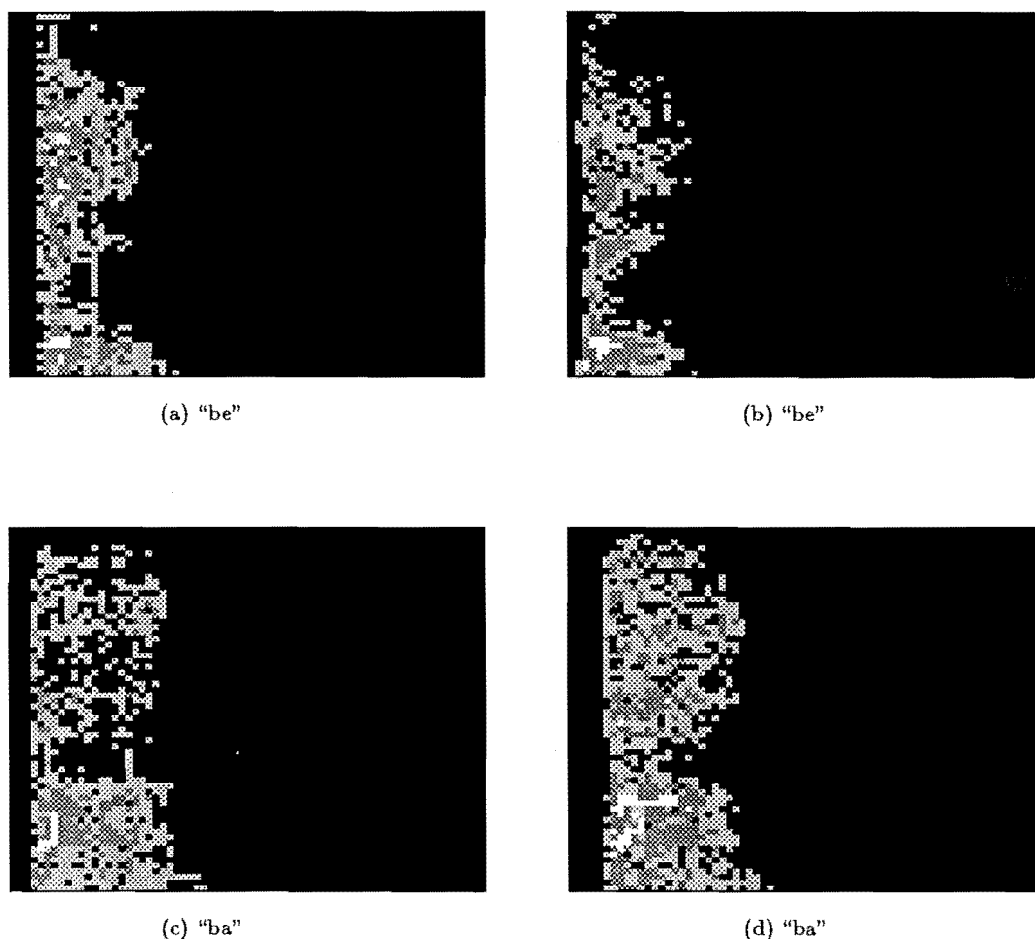


Figure 6.39. The spectral plots of target and error male utterances of "be" and "ba", which exemplify EE-13. The spectral plots of the target utterances are given in (a) and (b), the ones of the error utterances are given on the bottom two.

corresponding to the vowel [u], look fairly similar as is to be expected. The vowel [u] is the final sound in both "two" and "do".

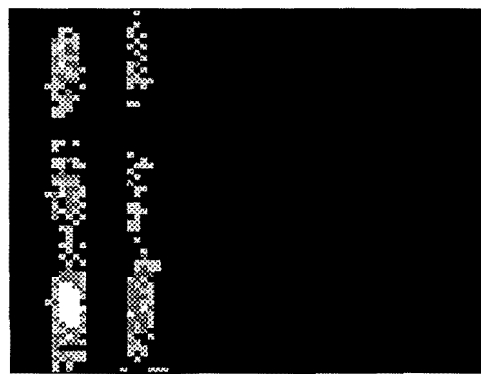
Figure 6.39 shows the plots of the male utterances of "be" and "ba" (the target and error utterances which exemplify EE-13). The utterances of "be" and "ba" differ in the final phoneme. It can be seen in fig 6.39 that the main difference between the target and error plots is the shape of the right half of the plots. All the utterances begin with the voiced plosive [b]. It can be seen that the left edges of all four plots look similar.

The spectral plots are too crude to obtain accurate information about the frequency of the formants. However when regions of high energy concentration differ quite dramatically for two sounds it is possible to distinguish between them, such as for [s] and [ʃ] in "see" and "she", as in fig 6.37.

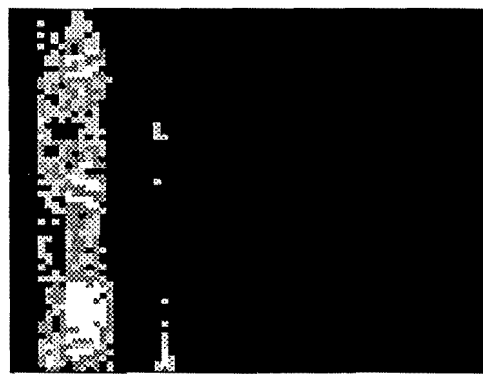
The width of the plots also played an important part in distinguishing between the target and error plots. It can be seen in figure 6.40 by comparing the widths of the first columns, that the syllable [pʊ] was more stressed in the female utterances of "pop" than in those of "poppa". In figure 6.41, the plots of the male utterances of "Ed" and "add", it can be seen that the [e] vowel is shorter in duration than the [æ] vowel.



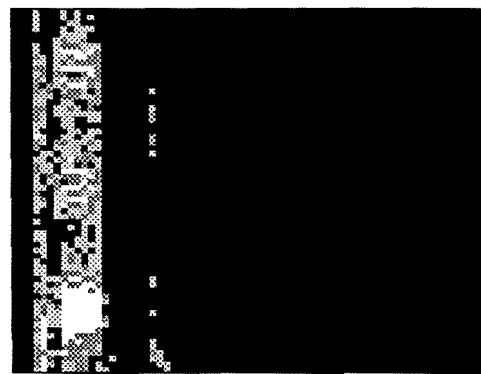
(a) "poppa"



(b) "poppa"



(c) "pop"

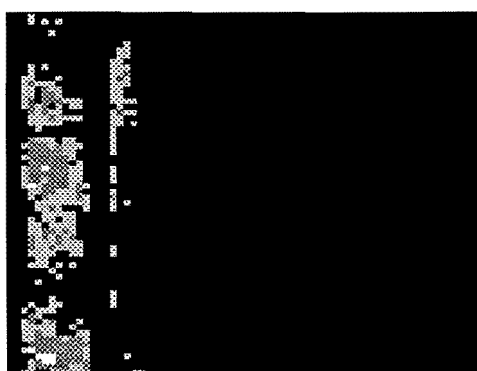


(d) "pop"

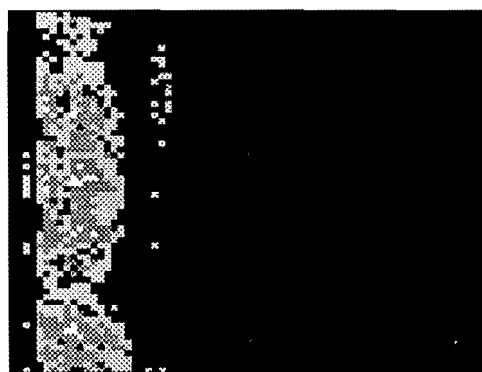
Figure 6.40. The spectral plots of target and error female utterances of "poppa" and "pop", which exemplify EE-4. The spectral plots of the target utterances are given in (a) and (b), the ones of the error utterances are given in (c) and (d).



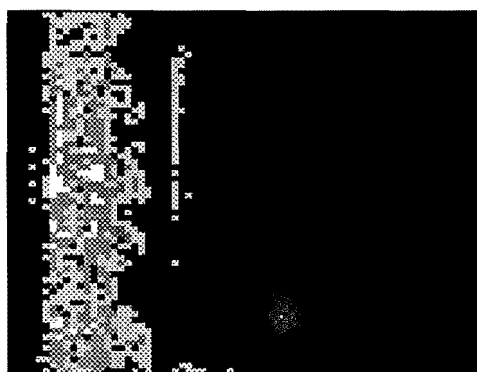
(a) "Ed"



(b) "Ed"



(c) "add"



(d) "add"

Figure 6.41. The spectral plots of target and error male utterances of "Ed" and "add", which exemplify EE-4. The spectral plots of the target utterances are given in (a) and (b), the ones of the error utterances are given on the bottom two.

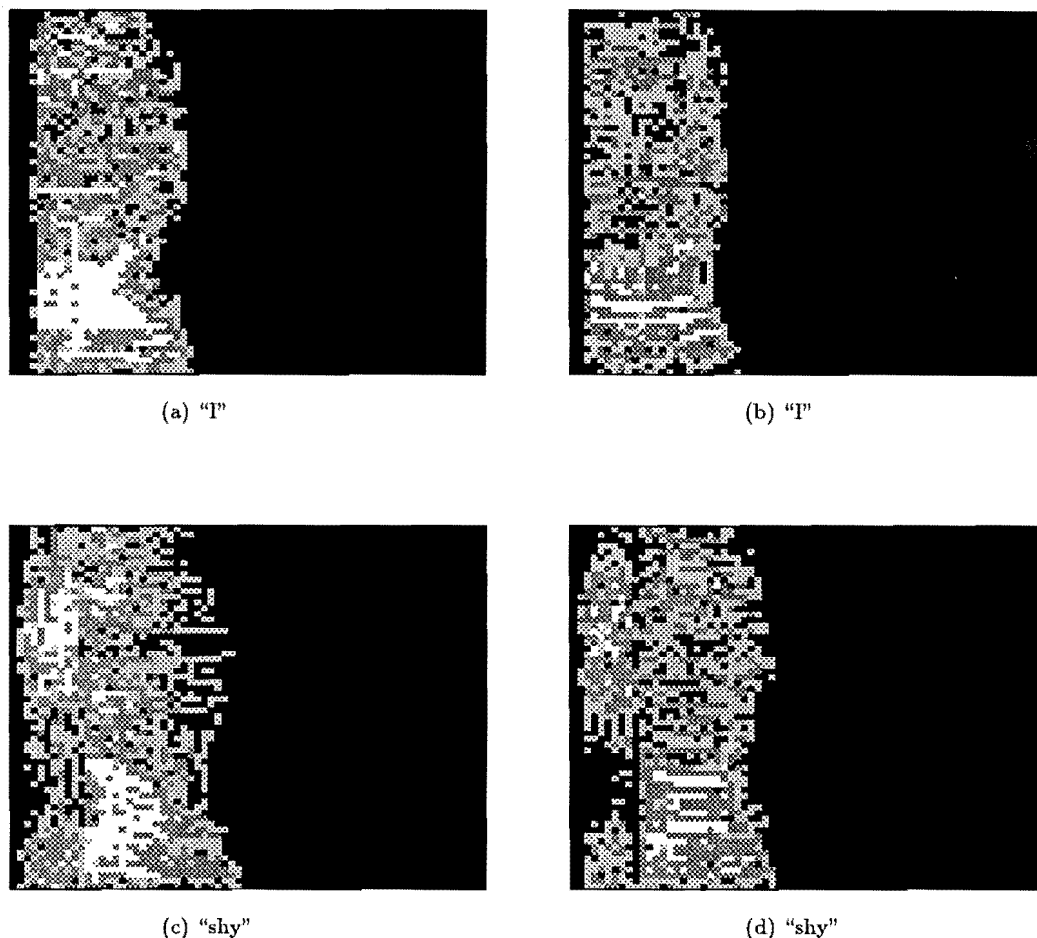


Figure 6.42. The spectral plots of target and error female utterances of "I" and "shy", which exemplify EE-6. The spectral plots of the target utterances are given in (a) and (b), the ones of the error utterances are given on the bottom two.

The colour distribution of the spectral shapes was determined by two factors, the frequency distribution of the sounds and the vocal intensity of the sounds. Thus it is possible that if a sound was uttered exceedingly loudly the entire plot would be white. The variation of vocal intensity between same-speech pairs caused a few distracting effects in some of the spectral plots. Two examples of this are given in figures 6.42 and 6.44.

Figure 6.42 gives the spectral plots of the female utterances of "I" and "Shy" and fig 6.43 gives the loudness contours of the same utterances. It can be seen in fig 6.43 that the first utterances of both "I" and "shy" were uttered much louder than the second utterances. The spectral plots of both the loud first utterances of "I" and "shy" contain far larger white regions than the plots of the second utterances. Thus it can be seen how vocal intensity affects the colour distribution.

In the spectral plots of the second utterances of "I" and "shy" there is a large dark grey region around the white regions. It is these regions that have turned white in the plots of the first utterances. The white region is when the frequency components are within 22 % of the maximum possible magnitude (256 levels). The dark grey regions

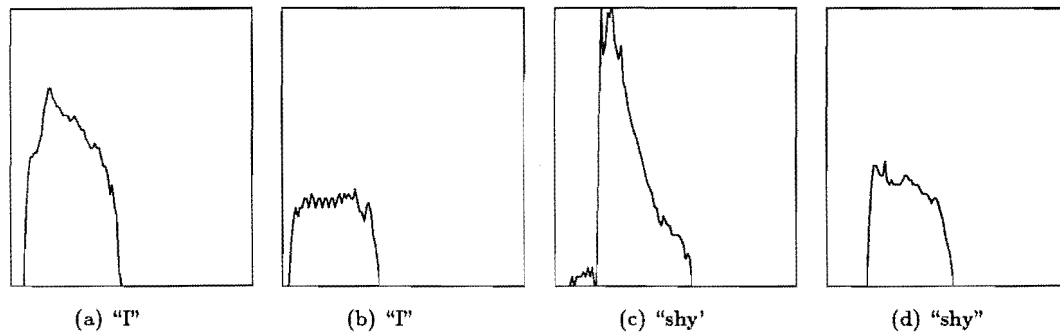


Figure 6.43. The loudness contours of target and error female utterances of “I” and “shy”, which exemplify EE-6. The contours of the target utterances are given in (a) and (b), the ones of the error utterances are given in (c) and (d).

(red in the original plots) are when the frequency components are between 58 % to 77 % of the maximum possible magnitude. The light grey regions (blue in the original plots) are when the frequency components are between 22 % and 57 % of the maximum possible magnitude and the black region is for the rest. Due to the coarseness of the magnitude quantization it is possible to see how it would be possible to have a large change in the magnitude of a frequency component and there would be no change in the colour which represents the magnitude. But conversely there could be a small change in the magnitude, which would result in a colour change.

By being aware of the effects loudness variations have on spectral plots it can be seen that the target and error plots in fig 6.42 of the utterances “I” and “shy” differ in the initial phoneme. The plots of “I” and the second half of the plots of “shy”, ignoring the effects of loudness variation, are fairly similar due to the diphthong [ai]. It can be seen that the left edge of the target plots is different from those of the error plots. The high energy concentration in the 2-4 kHz region for [ʃ] in “shy” can be seen by the white region in the upper left region of the error plots.

Figure 6.44 is another example where loudness variation has caused a distracting effect on the plots. It gives the plots of the male utterance of “fought” and “foot”. It can be seen by the widths of the first column in these plots that the vowel [ɔ] in “fought” is longer in duration than the vowel [u]. However the colour distribution of the two error plots is slightly different. There is very little white in the second error plot and a lot of black. This is because the second utterance of the “foot” was much softer than the first utterance, see fig 6.45.

It was also possible to distinguish suprasegmental aspects of speech on the spectral plots. The change in which portion of “now?” and “now!” is intensified (the target and error utterances which exemplify EE-21) could be seen in the spectral plots by the change in the colour distribution, as shown in figure 6.46. The utterance “now?” is more intense at the end of the utterance, whereas “now!” is more intense at the beginning. This can be seen by the shift in the white region from the right edge of the lower region to the left edge. The difference between “contract” (noun) and “contract” (verb) could be seen by the width of the plots (these are not shown in a figure). The colour distribution was fairly consistent across all four plots.

For all the elementary errors for which the spectral plots had remedial potential

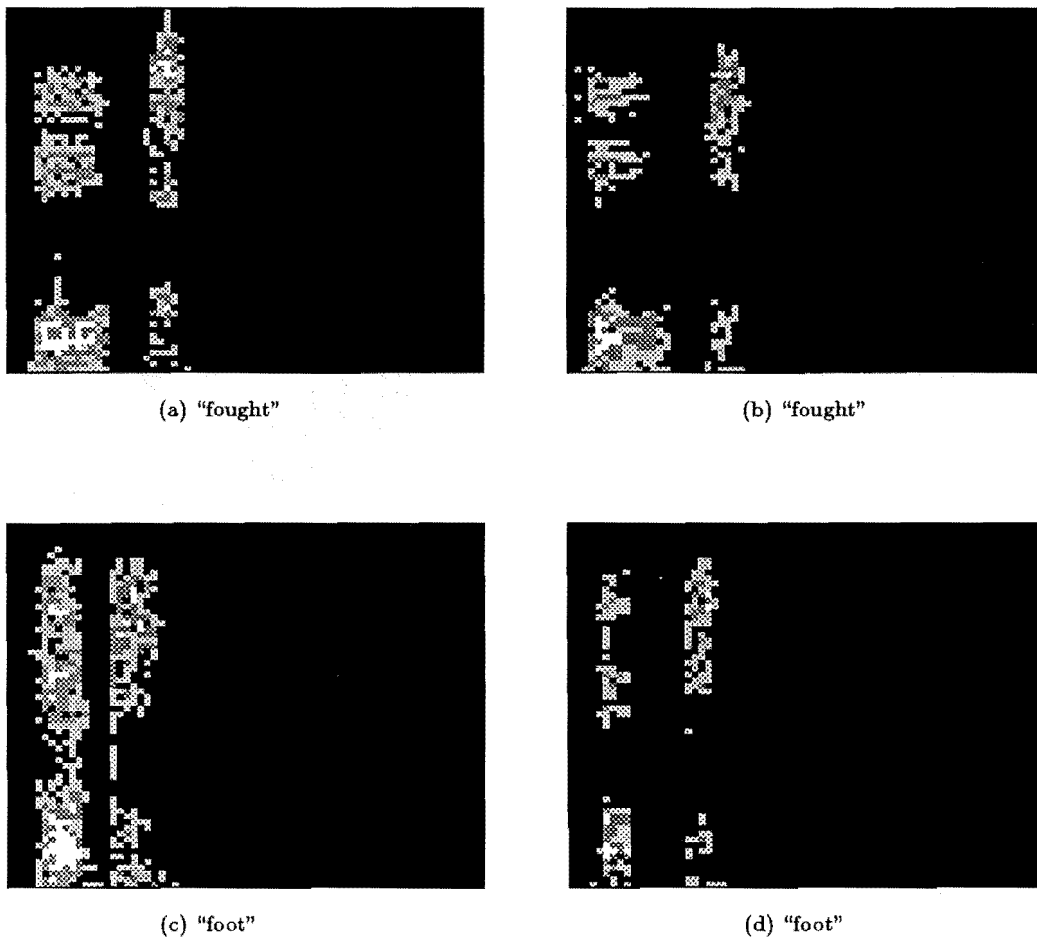


Figure 6.44. The spectral plots of target and error male utterances of "fought" and "foot", which exemplify EE-16. The spectral plots of the target utterances are given on the top two display graphs, the ones of the error utterances are given in (c) and (d).

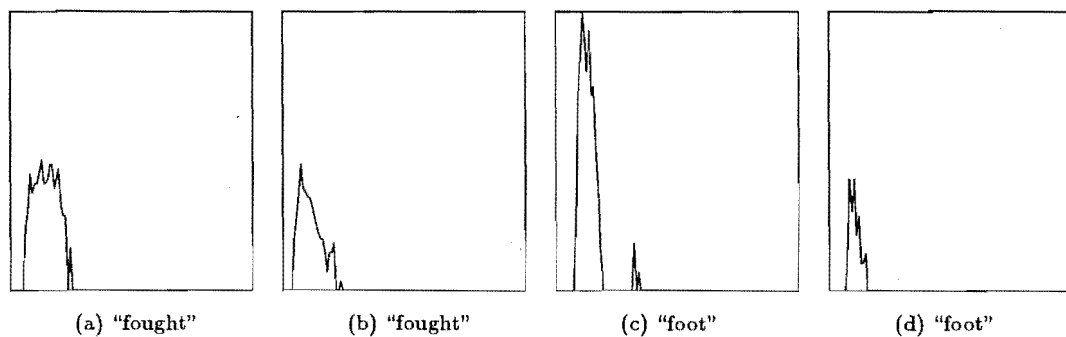


Figure 6.45. The loudness contours of target and error male utterances of "fought" and "foot", which exemplify EE-16. The contours of the target utterances are given on the top two display graphs, the ones of the error utterances are given in (c) and (d).

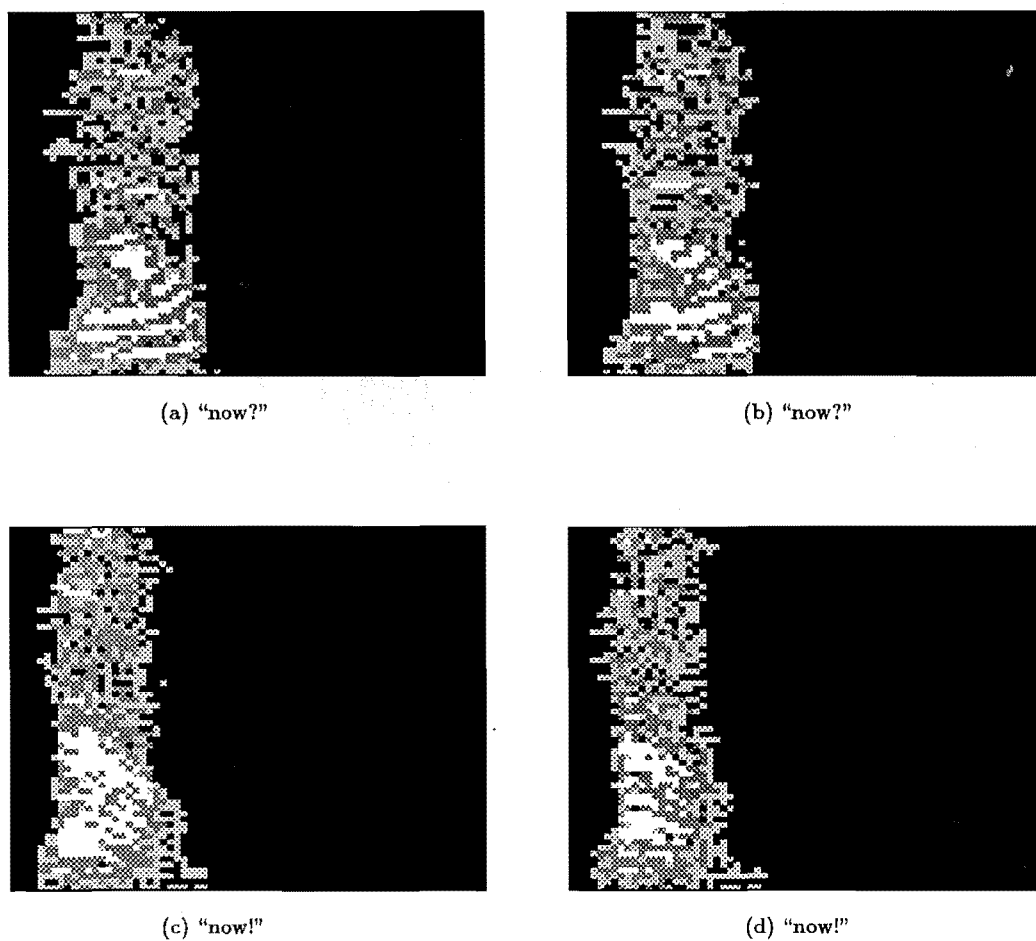


Figure 6.46. The spectral plots of target and error female utterances of “now?” and “now!”, which exemplify EE-21. The contours of the target utterances are given in (a) and (b), the ones of the error utterances are given in (c) and (d).

in VDT part I, the same plots also had remedial potential in VDT part II, as was found for the other time-plot display types. However in the VDT part II the spectral plots had remedial potential for an additional 6 elementary errors exemplified by female utterances, 14 compared with the 8 (out of 20) elementary errors in VDT part I (see table 6.11). For the elementary errors exemplified by male utterances the spectral plots had remedial potential for 13 additional elementary errors in VDT part II, 15 compared with the 2 (out of 20) elementary errors in VDT part I.

By comparing the results in VDT part I and VDT part II it was found that the most distracting feature in the spectral plots is their sensitivity to loudness variation. An investigation needs to be carried out as to how this can be avoided. The variations in loudness would have been due to one utterance being uttered more loudly than the other or due to the distance the microphone was from the mouth. The further away the microphone is from the lips the less intensity the sound will appear to have. This dependence on the distance of the microphone from the lips would also affect the loudness contours. This problem can be solved by ensuring the microphone is always at a fixed distance from the mouth, for example by attaching it to a head set. It is not so obvious how to lessen the effects on the spectrogram of utterances unintentionally uttered at different intensity if they are to be plotted in real-time.

Aside from the sensitivity to loudness, the VDT part II revealed that the spectral plot display type is very useful. It was difficult to know what to expect from the Spectrogram module from the comments made by the therapists in sections 5.2.1 and 5.3. Two of the therapists used the module effectively as an articulation corrector, while another felt the module imparted no useful information at all. The results of the VDT part II suggest that the Spectrogram module would be a useful articulation corrector; therefore it is highly likely that the therapists who used the module effectively were using the information from the screen to explain to their clients what was wrong with the clients' utterances. It is interesting to note that not many of the therapists who had the CASTT in their clinics for the short-term evaluations (two to three weeks) found the Spectrogram module (which has the spectral plots) very useful (see sec 5.1.2.1). However two of the three therapists who had used the module in the long-term evaluations (two to three months) found it to be one of the most useful in the CASTT (see sec 5.2.1). Clearly the more time the therapists had to learn how to interpret the displays the more useful the module became.

6.5.2 The Current-Value-Plots

The VDT part II revealed that the time-plots were very useful modules. At least one of the time-plot display types had remedial potential for the vast majority of elementary errors. The time-plots were the loudness contours, the pitch contours and the spectral plots. The current-value-plots did not fare so well in VDT part II. The test exposed the inadequacies of the current-value-plots and some limitations of the test itself. In sec 6.3.1 we discussed the elementary errors that could be tested on the current-value-plots (see table 6.3). On reflection four more elementary errors have to be removed from the table, as will now be explained.

The phonetic realization of the phonemes [d], [b], [j] and [w] (the target or error utterances associated with EE-3, EE-10, EE-18 and EE-19 respectively) cannot be uttered in sustained isolation. Any attempt to sustain these sounds results in sustaining

a vowel sound, most probably [a], rather than a consonant sound. Even when the above phonetic realizations are produced as short sounds a short vowel follows the consonant. For instance, if a person were asked to pronounce [b] they would most probably pronounce [bɪ]. This has consequences when testing the current-value-plots of the CASTT.

The displays of the current-value-plots used in the VDT are in fact snapshots of one instant of the display of an utterance. The snap-shot for sustained sounds would be fairly similar regardless of where in the duration of the sound it was taken ([ʃ] and [a] are examples of sustained sounds). This is not so for the phonetic realizations of the phonemes [b], [d], [w] and [j]. Any snap-shot from the current-value-plot will be showing some arbitrary portion of the transition from the consonant sound to the vowel sound.

When the target and error combinations of NZ-SL1 which exemplify EE-3, EE-10, EE-18 and EE-19 are reduced to phonetic realizations of phonemes they become [t/d], [m/b], [j,u], [w/r]. None of these combinations can be used to test the current-value-plots since [d], [b], [j] and [w] cannot be produced in sustained isolation and they are always followed by a vowel. The descriptions of the elementary errors are such that it is not possible to substitute [d], [b], [j] and [w] for sustained sounds. Thus the current-value-plots cannot be tested for remedial potential for those errors, nor can they have remedial potential for them.

The Lissajous figures are continually updated. The pattern remains fairly similar throughout the duration of a sustained isolated sound (such as [s] or [i]). The pattern gradually changes, however, for sounds like [b], [d], [w] and [j], as the sound goes through its transition into a vowel. There is no distinct moment when a characteristic pattern for any of those sounds would be seen. This also applies for the vocal tract module display (ignoring for the moment that the module was designed only to be used for vowel production). It would not be possible to see the intended difference between the target and error utterances. Therefore neither the Lissajous Figures nor the Vocal Tract Shape modules can have remedial potential for the above mentioned elementary errors.

The Fricative monitor also cannot be used to remediate EE-3, EE-10, EE-18 and EE-19. However, it is for a different reason than that given for the Lissajous figure and Vocal Tract Shape modules. The Fricative Monitor is meant to be an indication of frication. None of the target and error utterances which exemplify the above mentioned elementary errors are fricatives. Therefore theoretically the module should not provide a response for any of the target and error utterances.

The results of the VDT part II for the Vocal Tract Shape, Lissajous figure and Fricative monitor display types will now be discussed.

6.5.2.1 The Vocal Tract Shape Display

The Vocal Tract Shape module did not have remedial potential for the same elementary errors in VDT part II as it did in VDT part I. This finding was a surprise and it adds greatly to the importance of the VDT part II. In VDT part I the module had remedial potential for EE-14 exemplified by female utterances and EE-6, EE-8 and EE-10 exemplified by male utterances.

The elementary error EE-10 was not tested in VDT part II because of the module's

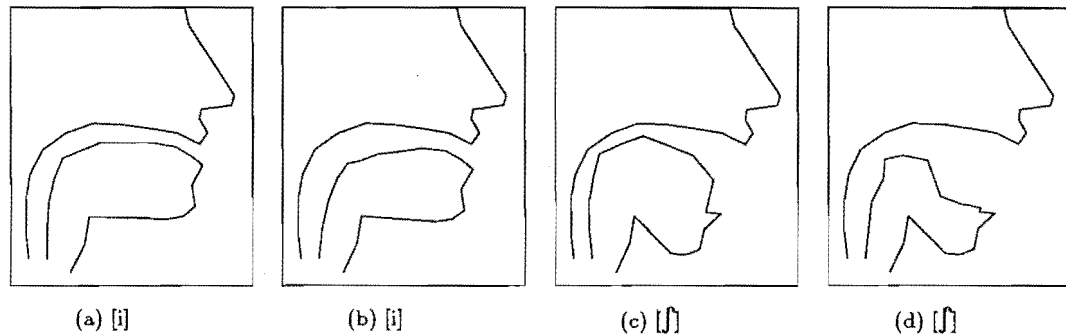


Figure 6.47. The Vocal Tract Shape Reconstructions of target and error male utterances of [i] and [ʃ] which exemplify EE-6. The reconstructions of the target utterances are given in (a) and (b), the ones of the error utterances are given in (c) and (d).

inability to display the short sound [b], ([m] and [b] are the target and error utterances which exemplify EE-10). The vocal tract shapes of the target and/or the error utterances which exemplify the three other errors the module had remedial potential for in VDT part I were incorrect. One example of this is the vocal tract reconstructions of the male target and error utterances [i] and [ʃ] which exemplify EE-6, given in fig 6.47. The vocal tract reconstructions of [i] are quite good. The mouth opening at the lips is narrow and the tip of the tongue is near the top front of the mouth. However the reconstructions for the [ʃ] utterance were incorrect. When producing [ʃ] the opening of the mouth is narrowed and the blade of the tongue is close to the hard palate. Neither of these attributes can be seen in the reconstructions of [ʃ] in fig 6.47.

The reconstructions of target utterances should look more similar than the reconstructions of the target/error different speech pairs. Similarly the reconstructions of the error utterances should look more similar than the reconstructions of the target/error different speech pairs. This is essentially the conditions for remedial potential in VDT part I. However the reconstruction of the vocal tract for [ʃ] is wrong and there is no evidence of the intended acoustic difference between vocal tract reconstructions of [i] and [ʃ]. Thus the module has no remedial potential for EE-6 exemplified by male utterances in VDT part II. In addition since the most obvious difference between the displays of the target and error utterances was not related to the intended acoustic difference between the utterances the implicit assumption in VDT part I (that is that the most obvious difference between the target and error displays can be attributed to the acoustic difference between the target and error utterance) was incorrect for this elementary error. In fact the implicit assumption was incorrect for all the elementary errors for which the vocal tract reconstruction display type had remedial potential in VDT part I.

The vocal tract shape reconstruction display did have remedial potential for EE-13, exemplified by either female or male utterances of [i] and [a], in VDT part II. We unfortunately omitted to test the module for remedial potential for this error in the VDT part I. Figure 6.48 shows the vocal tract reconstructions of the female utterances of [i] and [a]. The reconstructions of the vocal tract shape for [i] look similar, as do the reconstructions of the vocal tract for [a]. In the production of the sounds [i] and [a] the mouth opening is wider for [a] utterances than [i] utterances. In addition the

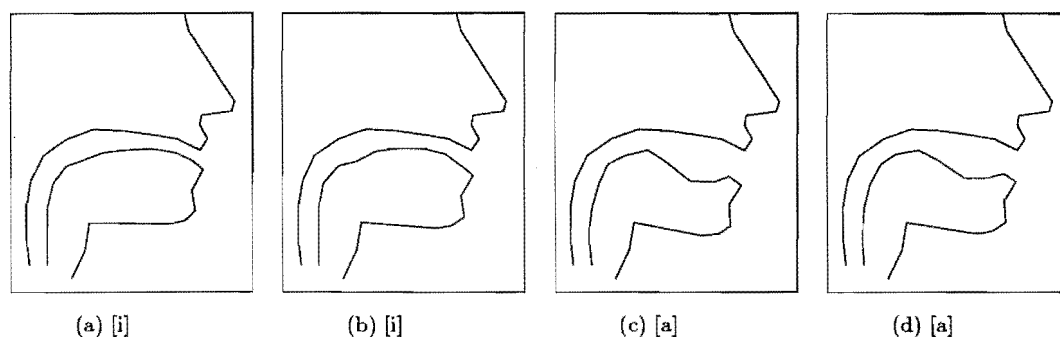


Figure 6.48. The Vocal Tract Shape Reconstructions of target and error female utterances of [i] and [a] which exemplify EE-13. The reconstructions of the target utterances are given in (a) and (b), the ones of the error utterances are given in (c) and (d).

tongue front is high for [i] and low for [a]. Both these attributes could be seen in the reconstruction of the vocal tract shape of the female utterances (see fig 6.48) and male utterances of [i] and [a].

The results of the VDT part II indicate that the module has no use in remediating errors in back vowels ([u], [ʊ], [ɔ]), nor could it remediate errors between the near neighbour front vowels [e] and [æ]. However the module could distinguish between the far-neighbour front vowels [i] and [a] and has remedial potential for EE-13. It is interesting to note that vocal tract reconstructions from the male utterance of [e] are also reasonably good. The mouth opening of [e] is wider than in the reconstructions for [i] but narrower than for the reconstructions of [a]. This is quite correct.

It is very encouraging that the vocal tract reconstructions of both the female and male utterances of [i] are similar, as are the reconstructions of the female and male utterances of [a]. This indicates that for these vowels the reconstructions are correct. To establish the latter the vocal tract reconstructions from the utterances of many more speakers would have to be observed. Two of the speech therapists in sec 5.2.1 reacted quite negatively towards the Vocal Tract Shape module and questioned the accuracy of the reconstructions (though they did not say which sounds they tested the module with). The results of the VDT part II suggest that the vocal tract reconstruction module could possibly be used as an aid for vowel production. However the reconstruction algorithm would need to be improved in order that the module could correctly reconstruct the vocal tract shape for many more vowels than it currently can. This conclusion was the same as that drawn from the therapists' comments. The proposed changes to the vocal tract shape reconstruction algorithm are discussed in chapter 8.

6.5.2.2 The Lissajous Figure Displays

It is not a straightforward process to assess the Lissajous Figures displays for remedial potential in VDT part II. The figures do not display a specific acoustic feature, unlike all the other modules in the CASTT. In the VDT part I the module had had remedial potential for EE-8 and EE-9 and EE-10 exemplified by female utterances and EE-6, EE-10 and EE-12 exemplified by male utterances in VDT part II. The module could not have remedial potential for EE-10 in VDT part II as was discussed earlier.

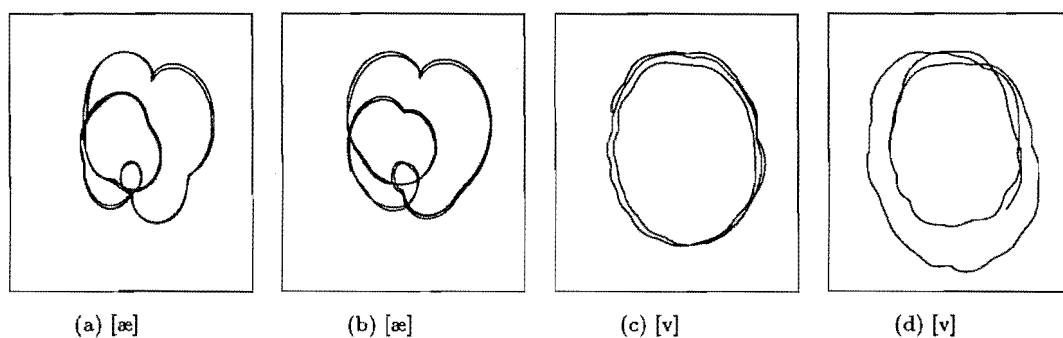


Figure 6.49. The Lissajous Figures of target and error female utterances of [æ] and [v] which exemplify EE-8. The figures of the target utterances are given in (a) and (b), the ones of the error utterances are given in (c) and (d).

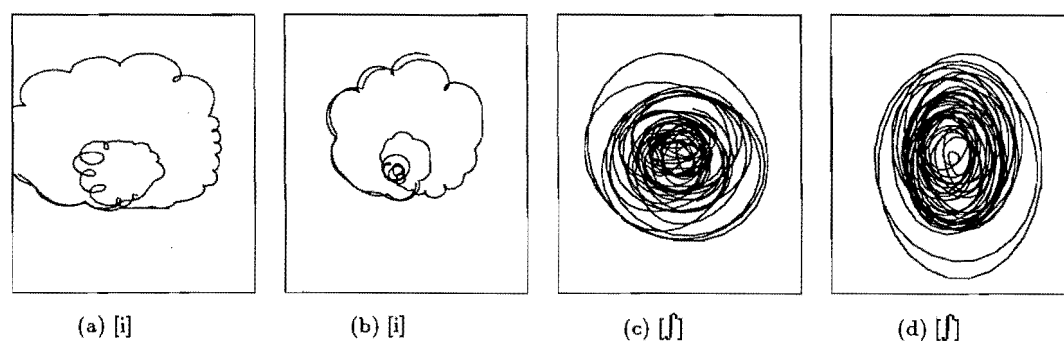


Figure 6.50. The Lissajous Figures of target and error male utterances of [i] and [ʃ] which exemplify EE-6. The figures of the target utterances are given in (a) and (b), the ones of the error utterances are given in (c) and (d).

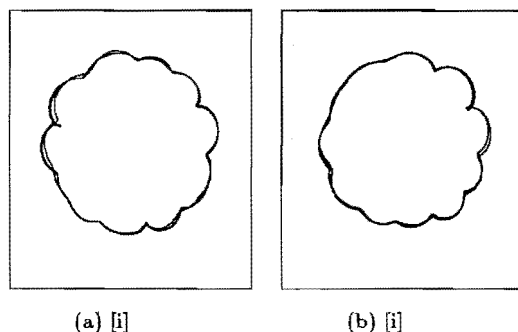


Figure 6.51. The Lissajous Figures of female utterances of [i].

The Lissajous figures for the utterances which exemplify the rest of the elementary errors for which the module had remedial potential in VDT part I conform to the first two criteria of remedial potential in VDT part II. Figure 6.49 shows the Lissajous figures of the female utterances of [æ] and [v], utterances which exemplify EE-8. Figure 6.50 gives the Lissajous figures of the male utterances of [i] and [ʃ], utterances which exemplify EE-6. The Lissajous figures of female utterances of [æ] look similar, as do the figures of the female utterances of [v], the figures of the male utterances of [i] and the figures of the male utterances of [ʃ].

It is questionable, however, how useful the Lissajous figures would be as an aid to articulation. The elementary errors for which the aid has remedial potential are different depending on whether the error was exemplified by female or male utterances. The pattern of isolated sounds is very dependent on the speaker. In no instances did the Lissajous figures for the female and male utterances of the same sound look similar. The Lissajous figures of the female utterances of [i] are given in fig 6.51 and those for the male utterances of [i] are given in fig 6.50 (a) and (b). It is true that the plots of the female utterances of [i] look similar as do the plots of the male utterances but the plots of female and male utterances of [i] are not similar. Thus a client producing a sound to match the therapist's pattern may not necessarily be producing the correct sound. There is also a lack of uniqueness in the figures for utterances of different sounds by the same speaker. The Lissajous figures for the female utterances of [v] and [z], for example, were very similar (see fig 6.52). For the reasons given above it was decided the Lissajous figures had remedial potential for none of the elementary errors exemplified in NZ-SL2.

The Lissajous Figure module has not been trialed by speech therapists as yet. The results of the VDT part II indicate that it is not an articulation aid. However, of all the current-value-plot display types in the CASTT at the moment, the Lissajous figures module could possibly be the most useful. The module could be used for practising phonation since it provided a visual response whilst a sound is phonated. The ability to phonate on command is something not every person can do. In addition the module has potential to be used for contrastive phonation in some instances as some different sounds produce different patterns. The Lissajous figures gave distinct repeatable patterns for the female utterances of [i], [ɔ], [æ], [s] and [m] and for the male utterances of [i], [ʃ],

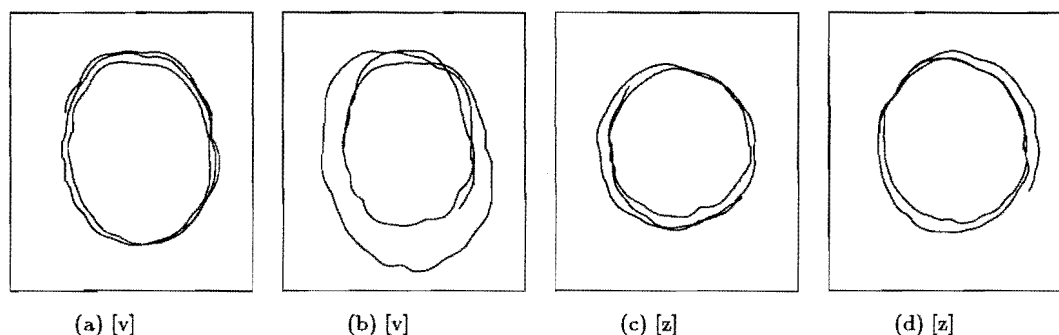


Figure 6.52. The Lissajous Figures of female utterances of [v] and [z]. The figures of the [v] utterances are given in (a) and (b), the ones of the [z] utterances are given in (c) and (d).

[u], [m] and [s]. Thus it is still worth trialing this module in the speech therapy clinics as a potential phonation aid for the CASTT.

6.5.2.3 The Fricative Monitor Display

The Fricative Monitor, whilst faring very well in VDT part I, did not look very good in VDT part II. The last two of the conditions for remedial potential in VDT part II were that the presence or absence of the speech feature of interest must be obvious from the visual displays and that the visual difference between the target and error displays be related to the intended acoustic difference between the target and error utterances. The intention of the fricative monitor was to indicate whether frication occurred. If frication was detected a horizontal bar should appear on the screen, if not the screen should remain blank. Thus it is evidence of frication or lack of it that we are looking for in the displays of the fricative monitor.

There were four elementary errors in NZ-SL2 which involved frication EE-6 (exemplified by [i] and [f]), EE-8 (exemplified by [æ] and [v]), EE-9 (exemplified by [s] and [z]) and EE-12 (exemplified by [s] and [f]). We expect the responses for non-fricative sounds to be a blank screen. The module only consistently indicated frication for the female and male utterances of [s] and the male utterances of [f]. The module had remedial potential for EE-6 (see fig 6.53) and EE-12 exemplified by male utterances in VDT part II as it did in VDT part I. It did not have remedial potential for EE-9 exemplified by female utterances, as it did in VDT part I as the module did not indicate frication for [z].

The inaccuracy of the module for displaying frication has long been a criticism by the speech therapists, as was the module's inconsistency in responding to frication (see sec 5.2.1). These too could be seen in the displays used in the VDT. The module indicated that only one of the two female utterances of [f] and of [v] were fricatives and only one of the two male utterances of [v] and of [z] were fricatives.

Another problem with the fricative monitor is its inconsistent response to a particular fricative uttered by two different speakers. Fig 6.54 gives the fricative content displays for the female and male utterances of [s]. The horizontal bar is longer for the male utterances than for the female utterances of [s]. A client producing a sound to match the therapist's pattern may not necessarily be producing the correct sound. This

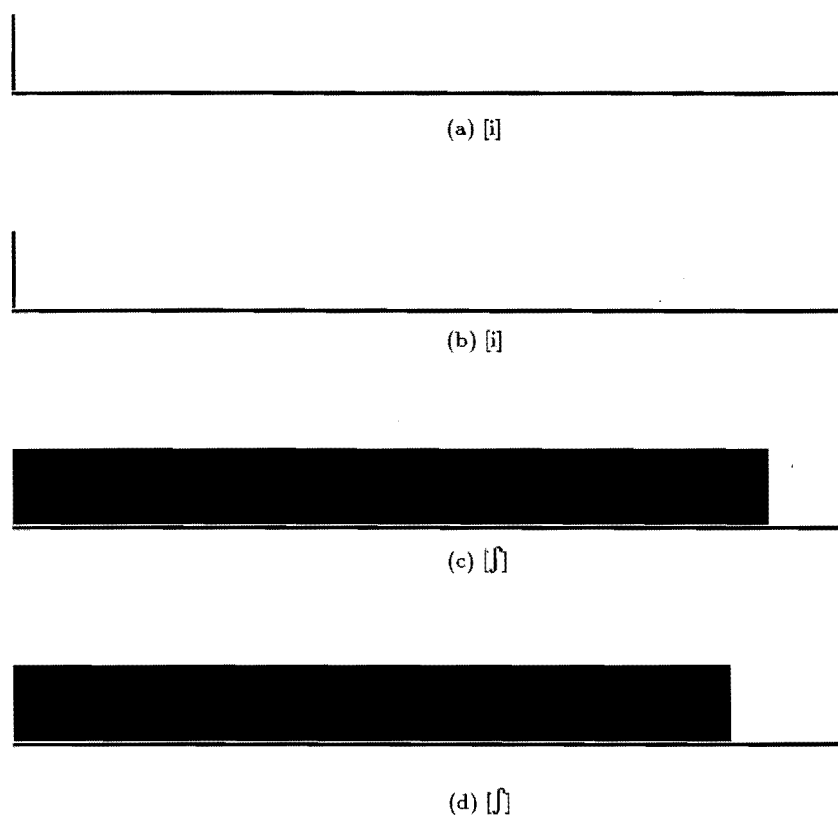


Figure 6.53. The Fricative Content response of target and error male utterances of [i] and [ʃ] which exemplify EE-6. The response to the target utterances are given in (a) and (b), the ones of the error utterances are given in (c) and (d).

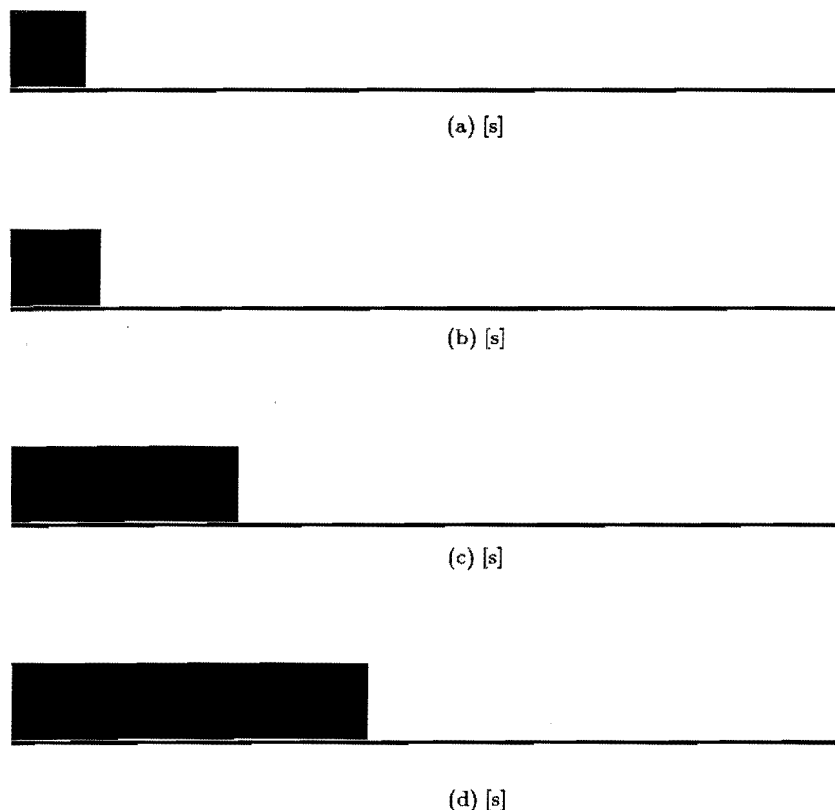


Figure 6.54. The Fricative Content response of the female and male utterances of [s]. The response to the female utterances are given in (a) and (b), the ones of the male utterances are given in (c) and (d).

is the same problem as that which would occur with the Lissajous figures.

At first glance it does not make sense that the zero-crossing rate (calculated for a 19.2 ms period, see sec 4.6.1) of the male utterances of [s] is greater than that of the female utterances (this is indicated by the length of the bar in fig 6.54). It is because all sounds are filtered at 4.5 kHz before being sampled at 10 kHz. The proportion of high frequency components with high magnitude is inversely proportional to the effective length of the vocal tract (Pentz *et al.*, 1979). Since women generally have shorter vocal tracts than men, women have a significant energy component in their fricative spectra at higher frequencies than men. Many of the frequency components for [s] are above 5 kHz, so the filtering removes this information. Thus the waveform of the male utterances of [s] has a greater zero-crossing rate than that of the female utterances. This highlights another problem with zero-crossing - its dependence on the pre-processing of the time domain waveform.

All the fricative sounds have a major high frequency component. The sounds [f,v,θ, ð, s, z] all have major frequency components above 5 kHz. These will all be removed by the low pass filter (which has a cut-off of 4.5 kHz on the SX8 board and a cut-off of 3.0 kHz on the TMSboard). This is possibly one of the reasons why the Fricative Monitor of the CASTT was so inconsistent. There is only a response on the screen if there is a significant energy component lower than 4.5 kHz in the fricative spectra, either due to the type of fricative being uttered or the effective length of the vocal tract of the speaker being long enough. One possible solution would be to increase the

sampling rate and low pass filter bandwidth. Whilst this would enable the monitor to be more responsive to frication it would not remove the speaker dependency problem. In addition the therapists also found the fricative monitor module too general. They did not want just an indication of frication; they wanted an indication that a specific fricative was being produced. The existing fricative monitor is not capable of doing this.

It was interesting to note that for both the female and male utterances of [a] the zero-crossing rates were greater than those for [i] ([a]/[i] exemplify EE-13) and the rates of the female and male utterances of [æ] were greater than those for [e] ([e]/[æ] exemplify EE-15). Ito and Donaldson (1971) found that zero-crossing rates could be used to differentiate vowel sounds for one speaker (a Canadian male). Whilst it is tempting to think the fricative monitor, or more specifically the zero-crossing technique, may have potential as an aid for vowel articulation, it does not. The non-uniqueness of the responses was a problem, as it was for fricative sounds. For example, the lengths of the horizontal bar for female utterances of [a] and [æ] were similar.

The elementary errors in the voiced/unvoiced speech set were not well serviced by the speech modules in the CASTT. The time-plots had remedial potential for EE-6 and EE-7 and the fricative monitor had remedial potential for EE-6. This is clearly an area where the CASTT could be improved. A new fricative monitor must be built and one that incorporates some recognition algorithm that can recognise each of the eight English fricatives.

6.5.3 Discussion

The VDT part II revealed that the time-plots (the loudness contours, the pitch contours and the spectral plots) had potential to be very good speech aids. Table 6.12 gives all the elementary errors, exemplified by female utterances and by male utterances, for which at least one of the time-plots had remedial potential in VDT part II, (they are indicated by a tick). For comparison the elementary errors for which the time plots had remedial potential in VDT part I are also given. It can be seen in VDT part II that the time-plots had remedial potential for 24 (out of a possible 29) elementary errors exemplified by female utterances and 23 (out of a possible 29) elementary errors exemplified by male utterances.

The results of the VDT part II established that for all but one instance when a time-plot had remedial potential for an elementary error in VDT part II it meant that the most obvious difference between the displays of the target and error utterances was related to the intended acoustic difference between the utterances. This means the results of the VDT part I for the time-plots can now be seen in an entirely new light. The results suggests that a speech therapy client, with no knowledge of phonetics or phonology would be able to use the displays of the CASTT to distinguish between the target and error utterances which exemplify a sizeable portion of the elementary errors. This could be done purely on the basis of judging which displays were the most similar. There is another and much larger set of elementary errors for which the aid has remedial potential if one has knowledge of what features to look for in the displays. A speech therapist would know the intended acoustic differences between target and error utterances and would know what features to look for in the displays. They could then relay this to the client.

| | | ART. INT. | | | | | UN./V. | | | | NAS. | | ART. SUB. | | | | | | | | SUPR. | | | SP. TIMB. | | | | | | |
|--------|-----------|-----------|---|---|---|---|--------|---|---|---|------|----|-----------|----|----|----|----|----|----|----|-------|----|----|-----------|----|----|----|----|----|----|
| TEST | UTTERANCE | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
| VDT II | Female | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| VDT II | Male | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | | | ✓ | ✓ | | ✓ | | ✓ | ✓ | ✓ | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| VDT I | Female | ✓ | | | ✓ | ✓ | | ✓ | | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | | ✓ | | | ✓ | | | ✓ | | | ✓ | | ✓ |
| VDT I | Male | ✓ | | | ✓ | | | | | | ✓ | ✓ | | ✓ | | ✓ | | | | | ✓ | | | | | | | | | |

Table 6.12. The elementary errors exemplified by female utterances and by male utterances for which at least one of the time-plots had remedial potential in VDT part II. For comparison the results of VDT part I have been included.

The VDT part II revealed that the current-value-plots (the vocal tract shape reconstruction, Lissajous figures and fricative content plots) were not good speech modules. The vocal tract shape reconstructions have remedial potential for EE-13 exemplified by both female and male speakers. The fricative content plots only have remedial potential for EE-16 and EE-12 exemplified by male utterances. The Lissajous figures module has remedial potential for no elementary errors in VDT part II. Ironically, of the existing current-value-plot modules, it probably has the most potential as a speech aid, as an aid for phonation. The 29 elementary errors did not test for anything as fundamental as the ability to show phonation. The patterns produced by the Lissajous figures when sounds are phonated would be engaging, especially for the young speech therapy clients.

The CASTT is quite weak on the elementary errors in the unvoiced/voiced set (see table 6.12). For this reason it was decided to build a new Fricative module. This will be discussed in detail in chapter 7. It was decided to see whether the Vocal Tract Reconstruction module could be improved since it correctly reconstructed the vocal tract shape for the female and male utterances of [i] and [a]. Chapter 8 discusses in detail the proposed improvements of the vocal tract reconstruction algorithm.

Both the VDT part I and part II are necessary to assess the CASTT. The VDT part II shows the elementary errors for which the CASTT has remedial potential if one knows what features should be evident in the displays of the target and error utterances. A speech therapist could carry out a similar test, providing they could get two displays from the target utterances and two from the error utterance on the screen simultaneously. In doing the test they would be able to assess how many elementary errors the aid has remedial potential for. The results of VDT part I for time-plots gives the elementary errors in which the difference between the displays of the target and error utterances is obvious. No expert knowledge is needed to distinguish between the target and error displays. The difference between the VDT part I and VDT part II results gives an indication of the displays in the CASTT which could be improved. By comparing the results of VDT part I and VDT part II for the loudness contours, for example, we found that the heights of the contours had to be made a more obvious feature. This would not have been revealed in VDT part II. Thus the results of VDT part I in conjunction with the results from part II are useful from a design point of view. They show where the displays could be improved.

There are two very important results to conclude from this chapter. Firstly, the VDT part I and part II are good tests to assess the remedial potential of a speech aid and to show where its displays could be improved. Secondly, the results of the VDT show that the current-value-plots of the CASTT do not have remedial potential for many elementary errors. However the VDT reveals that the time-plots of the CASTT

are all very effective modules. Whilst there are several improvements that need to be made to the displays, the time-plots have remedial potential for the vast majority of elementary errors.

CHAPTER 7

COMPUTER RECOGNITION OF FRICATIVE SOUNDS

7.1 INTRODUCTION

The previous chapter, chapter 6, assessed the visual displays of the CASTT by use of the VDT. The VDT indicated that whilst the CASTT had the potential to remediate a large number of speech errors it was not a good aid for speech errors involving frication (see sec 6.5.3). This means the current two modules in the CASTT intended for monitoring speech errors in frication, the Spectrogram and Fricative Monitor modules, are inadequate.

This chapter will outline the preliminary development of a new module in the CASTT; a new type of Fricative Monitor. The proposed monitor is to indicate if a pre-specified fricative has been articulated correctly. The motivation to build a new module did not come from the VDT, the VDT merely confirmed the need. It came from the speech therapists who evaluated the CASTT. The existing Fricative Monitor in the CASTT was not liked by many therapists in the evaluation. They felt it provided an inaccurate and inconsistent indication of frication (see sec 5.2.1). In addition the therapists felt that a module which only provided an indication of frication was of little use in speech therapy. They said the module should be an articulation corrector for each of the different fricatives in English.

Designing a new Fricative Monitor is not just a matter of developing a new algorithm on the existing CASTT. The actual hardware of the CASTT has to be changed if the CASTT is to be a speech aid for fricatives. The speech, processed by the TMS-board in the CASTT, is currently band-limited to 0 - 3000 kHz. This means most of the significant information in the fricative sounds is being removed, since fricatives have significant information in their spectrum up to at least 10 kHz (see sec 4.1). In sec 4.2 it was mentioned that provided one could, from the displays of the CASTT, distinguish between different sounds it did not matter if the sounds could not be accurately represented in CASTT modules. Whilst this comment still stands, in light of the results from the VDT, the bandwidth of the speech signal in the CASTT is obviously not enough to enable distinction between all the fricative sounds; what is more, so much information has been removed from the speech signal that not even a consistent indication of the presence of frication can be made. Therefore for the CASTT to be a speech aid for fricative sounds the sampling-rate and the bandwidth of the digitized speech must be increased.

The chapter will investigate the descriptors which can be used to characterise fricatives in English. It will then present several existing fricative recognition algorithms, including the findings of a preliminary investigation of a fricative recognition algorithm developed by myself and Brenda Satherley. Finally the relevance of these findings for

| | Labiodental | Dental | Alveolar | Palatal |
|----------|---------------|----------------|----------------|-------------------|
| Unvoiced | [f] as in fan | [θ] as in thin | [s] as in seal | [ʃ] as in shin |
| Voiced | [v] as in van | [ð] as in then | [z] as in zeal | [ʒ] as in leisure |

Table 7.1. The eight English fricatives with examples of the fricative sound in the initial position of the word, with the exception of the example for [ʒ] where it occurs in the medial position.

the proposed Fricative Monitor will be discussed. First, however, the phonetics of the fricatives in English will be briefly outlined.

7.2 PHONETICS OF FRICATIVES

Fricative sounds are produced by constriction in the vocal tract at some point in or above the larynx (Stevens, 1960). The constriction causes the air flow to shift from laminar to turbulent. Thus the constriction becomes a noise source (Badin, 1989). The shape of the vocal tract after the constriction acts as a filter (Fry, 1979). Hence the position of the constriction in the vocal tract strongly influences the fricative sound. Fricatives are either voiced or unvoiced. For unvoiced fricatives the only sound source is at the place of constriction. Voiced fricatives have two sound sources, the first is at the glottis and the second is at the place of constriction.

The eight English fricatives, listed according to place of articulation in unvoiced/voiced pairs in table 7.1, are: the labiodental fricatives [f,v] which are produced by a constriction between the lower lip and the upper teeth; the dental fricatives [θ,ð] which are produced by a constriction between the tongue and the upper teeth; the alveolar fricatives [s,z] which are produced by constriction between the tongue and the alveolar ridge; and finally the palatal fricatives [ʃ,ʒ] which are produced by a constriction between the tongue and hard palate. For each of the fricatives in table 7.1 an example of the fricative in the initial position of a word is given, with the exception of [ʒ] which only occurs in medial or final positions of words in the English language.

7.3 TIME DOMAIN FRICATIVE CLASSIFICATION

7.3.1 Classification due to Waveform Intensity

Stevens (1960), Jassem (1965), Stevens (1971), and Behrens and Blumstein (1988) have investigated the classification of fricatives analytically according to the relative intensity of the sound wave. Three of the four studies were only on voiceless fricatives (Stevens, 1960; Stevens, 1971; Behrens and Blumstein, 1988). The intensity of the sound wave was calculated from either the air pressure and air flow in fricative production (Stevens, 1960; Stevens, 1971) or by measuring the amplitude of the time domain wave form (Jassem, 1965; Behrens and Blumstein, 1988). The fricatives analysed were either extracted from isolated fricatives (Stevens, 1960) or from FV or FVC syllables, where F is a fricative, V is a vowel and C is a consonant other than a fricative (this nomenclature to indicate a fricative, a vowel or a nonfricative consonant will be used throughout this chapter) (Jassem, 1965; Behrens and Blumstein, 1988). Stevens (1971) does not mention how he obtained the fricative portions.

| | Duration Ranking From Shortest To Longest | | | | | |
|---|---|-------------|-------------|-------------|-------------|-------------|
| i | ð 112 ms | v 113 ms | θ 138 ms | f 159 ms | z 172 ms | s 182 ms |
| e | ð 77 ms | v 107 ms | f 135 ms | z 149 ms | θ 157 ms | s 165 ms |
| a | ð 97 ms | θ 106 ms | v 118 ms | f 137 ms | z 139 ms | s 175 ms |
| o | ð 106 ms | v 116 ms | z 139 ms | θ 140 ms | f 153 ms | s 166 ms |
| u | v 125 ms | θ 131 ms | ð 136 ms | f 59 ms | z 162 ms | s 181 ms |

Table 7.2. The average frication duration of three male speakers, for six of the English fricatives extracted from FV syllables, for five different vowels, based on data from Baum and Blumstein (1987).

All four studies state that, when considering unvoiced fricatives, the intensity of [f,θ] is always less than [s,f] (by about 10 to 15 dB (Jassem, 1965; Behrens and Blumstein, 1988; Jongman, 1989)). But distinctions between [f] and [θ] or between [s] and [f] are not really possible using intensity as a measure. The voiced fricatives were of greater intensity than unvoiced. Jassem found [ð,ʒ] to be 19dB greater than [f,θ], and [v,z] to be 21dB greater. Jongman(1989) stated that noise intensity is unlikely to be a major cue perceptually in the identification of fricatives; it is not a major cue analytically either. Interestingly McKinnon and Lee (1976) used sound intensity to distinguish [f,θ] from [s,f], when confusions between those sounds occurred.

7.3.2 Fricative Duration

de Manrique and Massone (1981) observed the fricative noise portions in FV FV syllables for Spanish fricatives. The syllables were obtained from four male speakers. It was found that the duration of frication was considerably less for voiced fricative than for unvoiced (de Manrique and Massone, 1981).

Baum and Blumstein (1987) observed the frication portion in FV syllables for the six English fricatives [f,v,θ,ð,s,z] and the American vowels [i,e,a,o,u]. The syllables were obtained from three male speakers. Baum and Blumstein found that for each vowel the duration of the fricative portion for unvoiced fricatives was always longer than their voiced counterpart. However the duration of frication varied according to the following vowel in the syllable, see table 7.2. There was also a considerable overlap between the duration distributions across all the fricatives and across all speakers (Baum and Blumstein, 1987).

When the fricatives are ranked in terms of frication duration for each of the five vowels, with the exception of [s] which was always the longest, no firm pattern emerges at first glance at table 7.2. But if the voiced and unvoiced fricatives are considered separately then from Baum and Blumstein's data it can be seen that [s] is always longer in duration than [f,θ] and [z] is usually longer than [v,ð].

These results suggest that at least for the five American vowels [i,e,a,o,u] that the voiced and unvoiced alveolar fricatives ([s,z]) have a longer fricative noise duration than their equivalent dental or labiodental counter parts. Frication duration is not going to serve as a primary cue for defining the place of articulation of a fricative. However

fricative duration may be useful as a secondary cue, for example to clarify confusions when [θ] utterances are incorrectly labelled as [s] utterances. It should be noted that from Baum and Blumstein's data it appears frication duration is dependent on the speaker.

7.3.3 Classification According To The Zero-Crossing Rate Of The Waveform

Ito and Robertson (1971) and McKinnon and Lee (1976) investigated classifying fricatives according to the zero-crossing Z rate of the time domain waveform. Z per interval of N samples was defined as:

$$Z = \sum_{n=1}^N [1 - \text{sgn}(x(n+1)) \text{sgn}(x(n))]/2 \quad (7.1)$$

and the zero-crossing rate of the first difference of the signal (Z') per interval of N samples was defined as:

$$Z' = \sum_{n=1}^N [1 - \text{sgn}(x(n+2) - x(n+1)) \text{sgn}(x(n+1) - x(n))]/2 \quad (7.2)$$

where $x(n)$ is the n th speech sample and $\text{sgn}(x(n))$ is the sign of that n th sample (Ito and Robertson, 1971).

The zero-crossing rate of a time domain signal is the rate at which the amplitude of the signal changes its sign. It has been described as a crude approximation to the spectral content in a signal (Rabiner and Schafer, 1978).

Ito and Robertson (1971) investigated the Z and Z' values calculated from 10ms portions of the unvoiced fricatives [f,s,ʃ]. The fricatives were extracted from FV syllables, where the vowels were [i,ε,u,ʌ,ɔ,æ] (spoken by an American speaker). All sounds were obtained from one male speaker, the speech being sampled at 16 kHz. Ito and Robertson showed that the unvoiced fricatives [f,s,ʃ] were distinguishable from each other on the basis of the Z and Z' values. McKinnon and Lee (1976) repeated Ito and Robertson's experiment on the fricatives [f,θ,s,ʃ]. Each of the Z and Z' values were obtained from 10ms portions of signal extracted from the fricative portion in FVC syllables, where C is a consonant other than a fricative. No mention is made of what vowels were used. The FVC syllables were obtained from the speech of one man sampled at 16.3 kHz.

Figure 7.1 illustrates the distribution of Z and Z' values for the unvoiced fricatives in $Z - Z'$ space calculated from McKinnon and Lee's research (1976). Both the mean, μ , and the standard deviation, σ , of all Z and Z' values for each fricative are shown. The numbers of fricative tokens from which the Z and Z' values were calculated, are given in the brackets. It can be seen that each of the four unvoiced syllables can be separately identified in $Z - Z'$ space.

7.4 FREQUENCY DOMAIN FRICATIVE CLASSIFICATION

The most successful descriptors for uniquely classifying each of the eight English fricatives have used the spectra of the fricative sound. In the research to be reviewed here the spectrum was calculated in various ways such as passing the signal through a filter

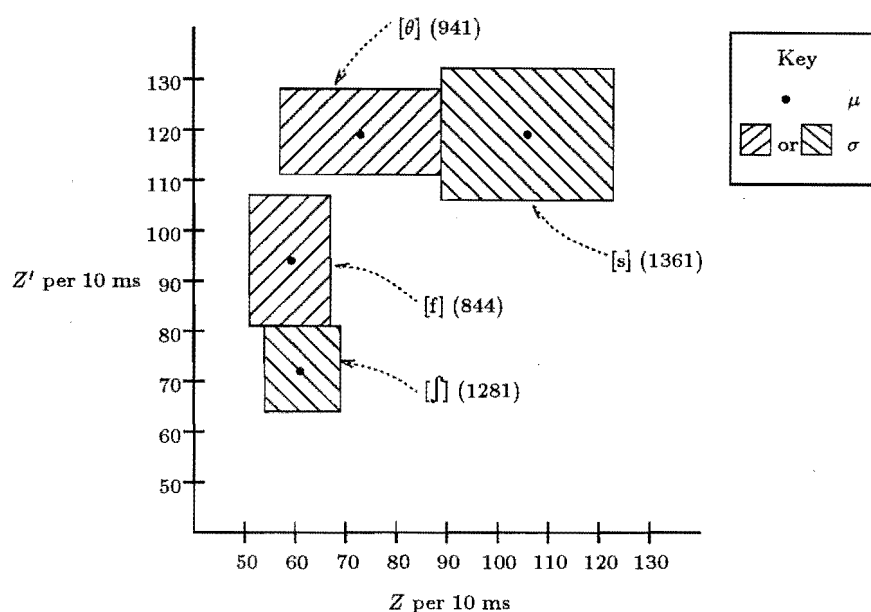


Figure 7.1. The separation of the four unvoiced fricatives in $Z - Z'$ space. The number in brackets is the number of fricative tokens from which the Z and Z' values were calculated (the data of this plot is from McKinnon and Lee (1976)).

bank, by performing a Fast Fourier Transform or linear predictive analysis on the signal. Table 7.3 gives some examples of the system or process used by various researchers to obtain the fricative spectra. In addition the table gives the spectral widths studied. It can be seen that the upper frequency limit on all the spectra is at least 9 kHz, which is a good deal greater than the 3 kHz in the spectra calculated in the Spectrogram module of the CASTT.

The fricative portions analysed in the literature were extracted from several different phonetic environments. Some chose to analyse fricatives spoken in isolation (Stevens, 1960). Most commonly the fricative portions were extracted from syllables or words where the fricative was in the initial position (Hughes and Halle, 1956; Jassem, 1965; Jassem, 1979; Molho, 1976; Pentz *et al.*, 1979; Behrens and Blumstein, 1988). The syllables and words used were usually selected so the vowels fol-

| Researcher | Means Of Spectral Analysis | Spectral Width |
|-------------------------|---|-----------------|
| Hughes and Halle (1956) | Hewlett Packard Wave Analyzer | 300 Hz - 10 kHz |
| Jassem (1965) | RASSLAN Spectrum Section Analyzer | 300 - 9 kHz |
| Molho (1976) | Linear Prediction Analysis on windowed speech | 600 - 9 kHz |
| Pentz (1979) | Kay Electric model 7029A | 60 Hz - 16 kHz |
| Baum (1988) | Linear Prediction Analysis | 0 - 9 kHz |
| Badin (1991) | 128 point Fast Fourier Transform on Windowed Speech | 0 - 10 kHz |

Table 7.3. Examples of the different methods used to obtained the spectra and the different frequency ranges various researchers have used in their investigation of spectral shapes of fricatives.

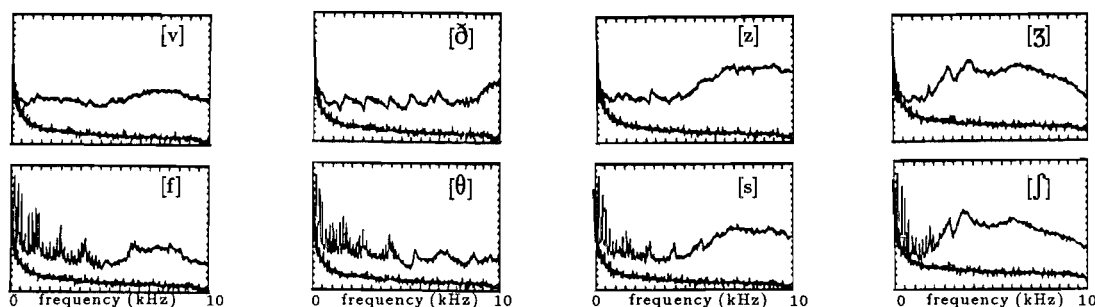


Figure 7.2. Typical spectra of the eight English fricatives [f,v,θ,ð,s,z,ʃ,ʒ] (the spectra are from (Badin, 1991) and are for an American female speaker).

lowing the fricative were an example of a front, centre and back vowel, for example Pentz *et al.* (1979) used the American vowels [i,æ,u]. Occasionally fricative portions extracted from medial and final positions in multisyllabic words were studied (Hughes and Halle, 1956; Jassem, 1979).

7.4.1 Fricative Spectral Shapes

The spectral shapes of the palatal [ʃ,ʒ], alveolar [s,z] and labiodental and dental [f,v,θ,ð] fricatives are readily identifiable (Hughes and Halle, 1956; Strevens, 1960; Jassem, 1965; Pentz *et al.*, 1979; Behrens and Blumstein, 1988). The typical shapes of the four pairs can be seen in figure 7.2. The spectra in figure 7.2 are taken from Badin (1991). They were calculated from the fricative portions extracted from FV syllables spoken by an American female speaker. Each plot is the average of 8 128 point FFTs, each computed on non-overlapped windowed speech samples (Badin, 1991). The top curve in each of the 8 plots in fig 7.2 is the spectrum of the fricative, the bottom curve being the average spectrum of the background noise of the room in which the fricative was recorded. The spectral shapes in figure 7.2 compare favourably with spectra obtained by Hughes and Halle (1956), Jassem (1965) and Soli (1981) and from the written descriptions about spectra given by Strevens (1960), Pentz *et al.* (1979) and Behrens and Blumstein (1988).

The first thing to observe from figure 7.2 is that the spectral shapes for each of the unvoiced/voiced fricative pairs do not differ considerably above 1000 Hz. The main difference between the unvoiced and voiced fricatives of each pair is a strong spectral component below 700 Hz for the voiced fricatives (Hughes and Halle, 1956) due to the vocal chords vibrating and to the harmonics of the pitch which can be seen in the 0-5 kHz region of the spectra of the voiced fricatives [v,ð,z,ʒ].

The second point to observe from figure 7.2 is the three distinct spectral shapes. The palatal fricatives, [ʃ,ʒ], have a major peak in their spectrum around the 2.5-3.5 kHz region. The spectral envelope rises rather steeply to this peak then drops away gradually. The alveolar fricatives, [s,z], have a major peak in their spectrum around the 4-6 kHz region. The spectral envelope rises gradually to this peak, drops away slightly then remains relatively flat. The labiodental and dental fricatives, [f,v,θ,ð],

| Researcher | Participants | θ | f | \int | s |
|-------------------------|----------------------------------|--------------------|--------------------|---------------|-------------|
| Hughes and Halle (1956) | 2 Men | - | - | 1.8 kHz | 3.7 kHz |
| Schwartz (1968) | 9 Men | - | - | 2.5 kHz | 5.5 kHz |
| Schwartz (1968) | 9 Women | - | - | 3.0 kHz | 6.5 kHz |
| Minifie (1973) | ? Men | 7-8 kHz | 6-7 kHz | 2.5 kHz | 4-7 kHz |
| Pentz (1979) | 21 Pre-adolescent Children | - | 11 kHz | 5.3 kHz | 8.4 kHz |
| Manrique (1981) | 4 Men | - | 1.5 - 8.5 kHz | 2.5 - 5 kHz | 5 - 8 kHz |
| Baum (1988) | 3 Men | 1.8-8.5 kHz | 1.8 - 8.5 kHz | 2.5 - 3.5 kHz | 3.5 - 5 kHz |
| Badin (1991) | 1 Man | relatively flat | relatively flat | 1.6 kHz | 4.8 kHz |
| Badin (1991) | 1 Woman | relatively flat | relatively flat | 3.5 kHz | 6.3 kHz |

Table 7.4. The positions of the major spectral peaks in the spectra of the English unvoiced fricatives. The data has been collected from studies by several different researchers.

have relatively flat spectra with no major peaks.

The small difference in the place of articulation between the labiodental and dental fricatives (constriction between the upper teeth and lower lip for the labiodental fricatives and constriction between the tongue and upper teeth for the dental fricatives) has been cited as the reason why the spectral shapes of the two fricative groups look so similar. However the fact that the sounds of [f,v] are perceptually distinguishable from [θ,ð] (Miller and Nicely, 1955) means that there are acoustic cues that distinguish [f,v] and [θ,ð] but as yet they have not been found.

The differences between the spectral shapes of the palatal, alveolar and labiodental and dental fricatives are obvious by visual comparison, as is the difference between the unvoiced and voiced fricatives of each of the fricative pairs. Most of the literature about the difference in spectral shapes of the various fricatives is descriptive, similar to that given above. Very little research has been done into turning these observed differences into quantifiable data. If a recognition algorithm is to be developed which recognises each of the eight English fricatives it is necessary to classify each of the fricatives uniquely. For this to happen a set of descriptors has to be found to satisfactorily characterise the differences in the fricative spectra of any English speaker, regardless of gender, age and phonetic environment.

Inspection of the spectral shapes of the four types of fricatives in figure 7.2 reveals that each spectrum has at least one major peak in it. Hughes' and Halle's (1956) research into the spectra of fricatives revealed that the major peaks in the spectra of [s,z] were at a consistently higher frequency than the major peaks of [ʃ,ʒ] and the spectra of [f,v] were observed to be basically flat up to 10 kHz (upper frequency limit of spectrum). Hughes and Halle did not investigate the spectra of [θ,ð].

Table 7.4 gives the position of the major peak in the spectrum of each of the four unvoiced fricatives in English, as found in several studies. Whilst the table gives the major peak position for the unvoiced fricatives only, the values can also be considered valid for the unvoiced part of the spectrum for each of the voiced fricatives from the respective unvoiced/voiced pairs.

Several generalizations about the fricative spectra of adult speech can be made from table 7.4. The major spectral peaks for [ʃ,ʒ] are typically within the 2.5 kHz to 3.5

kHz range and for [s,z] they are typically above 4kHz and less than 8kHz. The spectra of [f,v,θ,ð] are relatively flat up to 8 kHz. Whilst the spectral shapes of [s,z] [ʃ,ʒ] and [f,v,θ,ð] are distinctive, the distinction between [f,v] and [θ,ð] remains elusive (Baum and Blumstein, 1987).

Pentz *et al.* (1979) found the frequency bands where the major peaks occur in the fricative spectra of pre-adolescent children were much higher than those in the equivalent spectra for adults. In their study they found the only significant factor affecting the peak value was the effective length of the vocal tract, this is, the distance from the constriction (the place of articulation) to the lips. The position of the fricative in a word, the adjacent vowel, the vibrating or not of the vocal cords and the gender was not significant.

7.5 AUTOMATIC FRICATIVE SORTING ALGORITHMS

The last section presented several different methods which have been used to classify fricative sounds both in the time domain and in the frequency domain. This section will present several recognition (or sorting) algorithms which have been written to automatically recognise fricatives. These algorithms utilize the various characteristics of fricatives discussed in sec 7.3 and 7.4.

7.5.1 The McKinnon and Lee Zero-Crossing Rate Based Algorithm

McKinnon and Lee (1976) developed a recognition algorithm for the fricatives [f,θ,s,ʃ]. The algorithm used the zero-crossing rate Z of the signal and its first derivative Z' (see (7.1) and (7.2)) as well as the value of the maximum amplitude in the fricative portion, AMAX, and the zero-crossing rate of the signal after it has been passed through a 800-1900 Hz band-pass filter, Z_B .

The algorithm compared Z and Z' values, calculated from 10ms segments of unknown fricative sounds, to a statistical reference set and attempted to classify the fricative. The statistical reference set consisted of the mean \bar{Z} and \bar{Z}' values from a series of 10ms portions of the fricatives [f,θ,s,ʃ] and their respective standard deviations. The reference set data was obtained from the speech of one man only. This statistical reference set was plotted in fig 7.1.

A statistical distance measure (Q_i) in $Z - Z'$ space between each of the mean \bar{Z} and \bar{Z}' values and the Z and Z' of the unknown fricative was calculated, such that:

$$Q_i = \sum_{j=1}^2 \frac{1}{\sigma_{ij}} |x_j - \mu_{ij}| \quad (7.3)$$

where $i=1$ to 4 represents the i th class (i.e. one of the four unvoiced fricatives), $j=1$ to 2 represents the j th feature (i.e. either the Z or Z' values), x_j is the Z and Z' values of the unknown fricative and μ_{ij} and σ_{ij} are the mean and standard deviation of the Z or Z' values for fricative i . The unknown fricative was cursorily identified by finding the fricative i in the statistical reference set from which the smallest Q_i was calculated.

If the recognition process was stopped at this point McKinnon and Lee found they got recognition rates for the fricative sounds between 76.8 % and 90.6%. They found a significant number of [θ] were being identified as [s] using the distance measure classification and a significant number of [f] were being identified as either [θ] or [ʃ].

In order to overcome these confusions McKinnon and Lee did two things. Firstly they used the value of the maximum amplitude (AMAX) from the fricative portion and compared them with pre-stored values in a reference data set. Since the intensity of [s,f] is always greater than [f,θ] (Jassem, 1965; Behrens and Blumstein, 1988) the value AMAX in a fricative portion of [s] and [f] is always much higher than for [f] and [θ]. Thus the confusions in identifying [θ] as [s] or [f] as [f] were avoided.

Secondly, to resolve the confusion of [f] being identified as [θ] all the fricatives identified as [f] were filtered with an 800-1900 Hz bandpass filter and a new zero-crossing value, Z_B was calculated from the filtered signal. The new zero-crossing values Z_B for all the [θ] were found to be much lower than those for [f]. Thus the confusion of identifying [θ] as [f] was avoided. The recognition rates for the modified recognition algorithm ranged from 88 % to 98 % (see table 7.5).

7.5.2 Algorithms Based On Spectral Shape

Several methods have been employed to classify the fricative spectral shapes. Two studies divided the spectra up into a series of frequency bands (Hughes and Halle, 1956; Jassem, 1979). A crude estimate of the position of the major peak can be made by finding which of the bands holds the largest area under the spectral envelope and then identification of the fricative group can be made. Two other studies developed a recognition algorithm based on identifying the envelope of the spectrum (Molho, 1976; Jassem, 1979).

All the following discussions about recognition algorithms assume the spectrum of the fricative has already been calculated.

7.5.2.1 The Hughes and Halle Algorithm

The fricative sorting algorithm of Hughes and Halle (1956) separated the fricatives into one of three groups, the palatals, the alveolars or the labiodentals and dentals. It did not distinguish between the unvoiced and voiced fricatives in an unvoiced/voiced pair. The algorithm (see figure 7.3) calculated:

$$\Delta\mathcal{E}_{fsf} = \mathcal{E}_{(720Hz\ to\ 10kHz)} - \mathcal{E}_{(4.2kHz\ to\ 10kHz)} \quad (7.4)$$

$$\Delta\mathcal{E}_{fs} = \mathcal{E}_{(720Hz\ to\ 6.5kHz)} - \mathcal{E}_{(720Hz\ to\ 2150Hz)} \quad (7.5)$$

and

$$\Delta\mathcal{E}_{ff} = \mathcal{E}_{(max\ peak)} - \mathcal{E}_{(720Hz\ to\ 1370Hz)} \quad (7.6)$$

where $\mathcal{E}_{(a\ to\ b)}$ is the energy in dB in the band from a Hz and b Hz. No indication was given as to the value of $\mathcal{E}_{(max\ peak)}$. It was defined as “the energy in a 500Hz band centered at the peak frequency” (Hughes and Halle, 1956).

The first difference, $\Delta\mathcal{E}_{fsf}$, sorted the fricatives into one of two groups, [f,s] or [f,f]. If $\Delta\mathcal{E}_{fsf}$ was less than 2 dB the fricatives were labelled as [f,s] group otherwise they were labelled as the [f,f] group. Since the spectrum of [s] was observed to have a large energy component above 4kHz, $\Delta\mathcal{E}_{fsf}$ was expected to be much smaller for [s] utterances than [f] utterances.

The next difference $\Delta\mathcal{E}_{fs}$ identified the fricatives in the [f,s] group as either [s] or [f]. If $\Delta\mathcal{E}_{fs}$ was greater than 10dB the sound was classified as an [s]; if it was less than 5dB it was classified as an [f]. Since the spectrum of [f] was virtually flat compared to

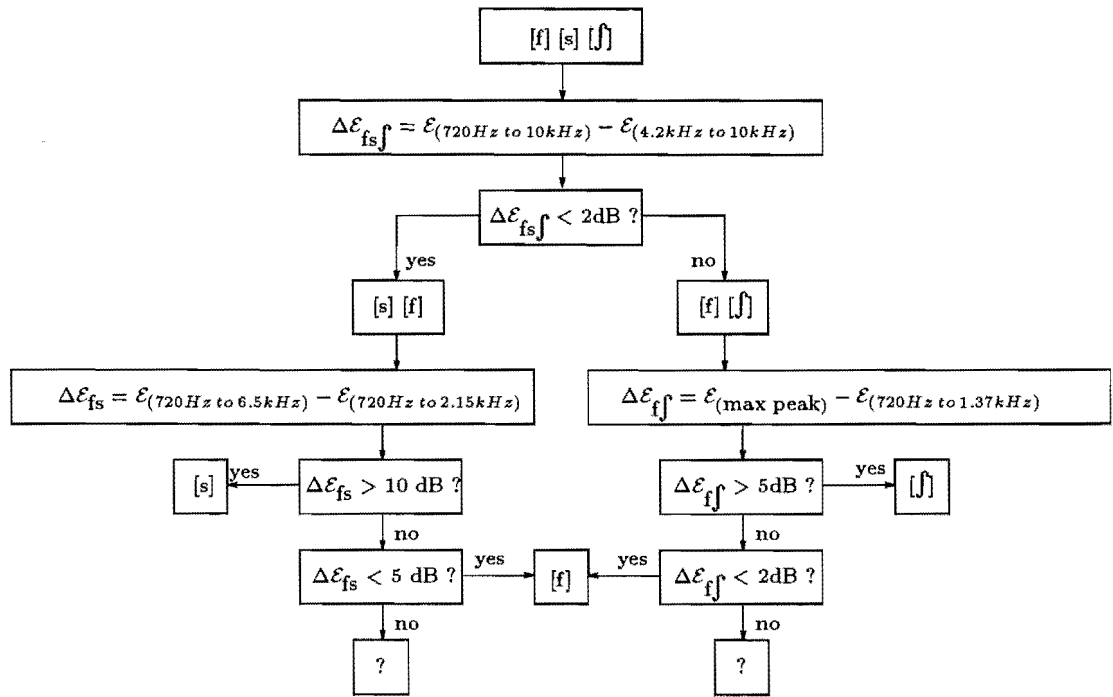


Figure 7.3. The fricative sorting algorithm proposed by Hughes and Halle (1956), where $E_{(a \text{ to } b)}$ is the energy in dB in the band from a Hz and b Hz.

the spectrum of [s], ΔE_{fs} is expected to be much smaller for [f] utterances than for [s] utterances. The last difference ΔE_{ff} identified the fricatives placed in the [f, ʃ] group as either [ʃ] or [f]. If ΔE_{ff} was greater than 5 dB the sound was classified as [ʃ]; if it was less than 2 dB the sound was classified as [f]. Due to the flatness of spectrum [f], ΔE_{ff} is expected to be much less for [f] than [ʃ]. No mention was made as to what happened in the algorithm if $5 \text{ dB} < \Delta E_{fs} < 10 \text{ dB}$ or if $2 \text{ dB} < \Delta E_{ff} < 5 \text{ dB}$. The recognition rates using Hughes and Halle's algorithm ranged from 77.4 % to 79.9 % (see table 7.5 for details).

7.5.2.2 The Jassem Algorithm

Jassem (1979) in his work with Polish fricatives [f, s, ʃ, ɣ, x, v, z, ʒ, ʒ̥] developed several fricative sorting algorithms. The first of these involves dividing the fricative spectrum into frequency bands and comparing the area in the bands. It is outlined in this section. The second algorithm involves approximating the spectral envelope with a polynomial. It will be outlined in the following section.

Jassem divided up the fricative spectrum into frequency bands and calculated the area in each region between the spectral envelope and a line drawn at 20 dB below the major peak. He called these areas partial areas. Jassem experimented with the number of bands the spectrum was divided into, trying from two to twelve bands. Using the area values obtained from each of the bands Jassem obtained inequality expressions which represented the spectral shape for each of the voiced/unvoiced pairs. For example Jassem found that if he divided the spectrum into three frequency bands (1 - 3.2 kHz, 3.2 - 5 kHz and 5 - 8 kHz) then the spectrum of the fricatives could be defined in the

following manner:

$$\begin{aligned}
 [f,v] \quad & A_1 > A_2 < A_3 < A_1 \\
 [s,z] \quad & A_1 < A_2 < A_3 > A_1 \\
 [ʃ,ʒ] \quad & A_1 < A_2 > A_3 < A_1 \\
 [ç,ʝ] \quad & A_1 < A_2 > A_3 > A_1 \\
 [x] \quad & A_1 > A_2 > A_3 < A_1
 \end{aligned}$$

where A_1, A_2, A_3 are the partial areas of the three frequency bands. The recognition rates using Jassem's frequency band method ranged from 27 % to 100%. It was found that the more bands the spectrum was divided into, the more accurate was the sorting algorithm, but the longer the algorithm's processing time - the ubiquitous trade off between accuracy and efficiency.

Table 7.5 gives the recognition rates Jassem got from dividing the spectrum up into three, four and twelve bands. Jassem felt that representing the spectrum with four frequency bands was the most effective computationally. It yielded reasonably high recognition scores (82 % to 99%) and the dimensionality was still relatively low (c.f. twelve bands).

7.5.2.3 Jassem's Envelope Polynomial

Jassem used curve-fitting techniques to obtain a fifth and sixth order polynomial which fitted the spectral envelope of each type of fricative (labiodental, alveolar, palatal, etc). He selected fifth and sixth order polynomials on the basis that they were of a high enough order to give a reasonable approximation to the envelope but were of low enough dimension to maintain computational efficiency (Jassem, 1979). A reference database was collected which contained sets of coefficients from polynomials which represented each of the spectral shapes of the fricatives. The set of polynomial coefficients representing the spectrum of the unknown fricative was compared with the sets of coefficients in the reference data base. The unknown fricative was identified as the fricative in the reference set from which the closest coefficient match was made. Using this method Jassem got recognition rates between 47 % and 100 %. Details of the recognition rates from using the fifth order polynomial are given in table 7.5.

7.5.2.4 Molho's Spectral Envelope Descriptors

Molho (1976) used seven descriptors in order to classify the spectral envelopes of the eight English fricatives. These were: (1) the lowest frequency peak f_{LP} ; (2) the second lowest frequency peak f_{SLP} ; (3) the sharpness, P_{sharp} , of the peak at f_{LP} calculated with

$$P_{sharp} = 10 \log(S(f_{LP})) - 5[\log(S(f_{LP} - 312.5)) + \log(S(f_{LP} + 312.5))] \quad (7.7)$$

where $S(x)$ is the magnitude of the spectrum at x Hz; (4) the spectral skew, which is the highest frequency, f_k , with a magnitude $S(f_k)$ such that :

$$S(f_k) \geq S_{ave} \quad (7.8)$$

where

$$S_{ave} = \frac{1}{n} \sum_{i=1}^n S(f_i) \quad (7.9)$$

and $S(f_i) \dots S(f_n)$ are the individual points in the magnitude spectrum; (5) the maximum spectral amplitude; (6) a boolean variable which is set true if $S(f_{LP})$ is within 6dB of the spectral maximum; and finally (7) a second boolean variable which is set true if $S(f_{SLP})$ is within 6dB of the spectral maximum.

No information was given in Molho's 1976 paper on how these descriptors were used to classify the spectra. However the above list does illustrate the types of descriptor that could be used for classification. The Molho's recognition rates using the above descriptors were between 57 % and 94% (see table 7.5 for details).

7.5.2.5 The Success Rates Of Several Fricative Sorting Algorithms

It is difficult to compare the success rates of the different fricative sorting algorithms due to the differing data sets the algorithms were designed for, the different methods of preparing the data and the differing spectral widths, as mentioned earlier in section 7.4. Thus no method can be declared any better than any other on the basis of the data given. In addition the groups of fricatives the algorithms recognised differed. Molho's algorithm recognised seven fricatives [f,v,θ,ð,s,z,ʃ,ʒ]. McKinnon and Lee's algorithm was designed to recognise each of the four unvoiced fricatives [f,θ,s,ʃ]. Hughes and Halle's algorithm recognised the fricatives as palatal, alveolar or labiodental/dental. The algorithm Jassem used was developed for Polish fricatives ([f,s,ʃ,ç,x,v,z,ʒ,ʒ̥]). Whilst there are small differences in the spectra of the English and Polish fricatives [f,v,s,z,ʃ,ʒ] (there are no [θ, ð] sounds in Polish) the overall shapes are fairly similar (Jassem, 1965). Therefore if Jassem's algorithms were used on the English fricatives [f,v,s,z,ʃ,ʒ] similar recognition rates could be expected.

However despite the fact that the recognition results came from quite different studies it is still interesting to observe the various success rates of the different algorithms. These are listed in table 7.5, along with the frequency width of the spectrum, the duration of the fricative portion analysed, what context the fricative portion was extracted from, the genders of the speakers and the number of speakers the algorithm was tested with.

7.6 A PRELIMINARY INVESTIGATION INTO A FRICATIVE RECOGNITION ALGORITHM

A Fricative recognition algorithm was developed by myself in collaboration with Brenda Satherley. Brenda was a final year student at the time, here at the department of Electrical and Electronic Engineering. Her final year project was the fricative recognition algorithm. The algorithm was designed to recognise each of the eight English fricatives. It was developed from data obtained from isolated fricatives. I then tested it with data obtained from fricative portions extracted from initial, medial and final positions in words. The algorithm developed here is similar to Jassem's fricative recognition algorithm using partial areas (see sec 7.5.2.2).

A data base of isolated fricative sounds was compiled from two female and three male speakers. Between two to five utterances of each fricative were obtained from each speaker. The utterances of the isolated fricatives were all sustained and were at least 0.25 seconds in duration.

The system used to record the sounds was similar to that illustrated in fig 4.1. A

| Researcher | Spectral Width | Speakers | Duration of Fricative Portion | Position In Word | [f] | [v] | [θ] | [ð] | [s] | [z] | [ʃ] | [ʒ] |
|--------------------------------------|----------------|----------------|-------------------------------|---------------------------------|--------------------|--------------------|-----|-----|--------------------|--------------------|--------------------|--------------------|
| Molho (1976) | 600-9000 Hz | 4 speakers | 10 ms | initial (FVC) | 88% | 57% | 69% | 57% | 94% | 73% | 83% | - |
| Hughes and Halle (1956) | 300-10,000 Hz | 3 men, 2 women | 50 ms | initial and final (FVC and CVF) | 78.6% ¹ | 78.6% ¹ | - | - | 79.8% ¹ | 79.8% ¹ | 77.4% ¹ | 77.4% ¹ |
| Jassem (1979) ² | 1000-8000 Hz | 3 men | 40 ms | initial and medial (FVFV) | 81% ³ | 71% ³ | - | - | 92% ³ | 94% ³ | 75% ³ | 73% ³ |
| Jassem (1979) ² | 1000-8000 Hz | 3 men | 40 ms | initial and medial (FVFV) | 95% ⁴ | 92% ⁴ | - | - | 90% ⁴ | 99% ⁴ | 82% ³ | 88% ⁴ |
| Jassem (1979) ² | 1000-8000 Hz | 3 men | 40 ms | initial and medial (FVFV) | 99% ⁵ | 100% ⁵ | - | - | 99% ⁵ | 100% ⁵ | 97% ⁵ | 100% ⁵ |
| Jassem (1979) ² | 1000-8000 Hz | 3 men | 40 ms | initial and medial (FVFV) | 88% ⁶ | 77% ⁶ | - | - | 96% ⁶ | 87% ⁶ | 88% ⁶ | 80% ⁶ |
| McKinnon and Lee (1976) ⁷ | 80-8000 Hz | 1 man | 10 ms | initial (FVC) | 77% | - | 77% | - | 91% | - | 90% | - |
| McKinnon and Lee (1976) ⁸ | 80-8000 Hz | 1 man | 10 ms | initial (FVC) | 88% | - | 94% | - | 96% | - | 98% | - |

notes

1

In Hughes and Halle (1956), the recognition algorithm identifies the Fricatives as either palatal, alveolar or labiodental groups (no distinction is made between labiodental and dentals) and it made no unvoiced/voiced distinction. The values quoted were obtained from the data presented in the paper as no actual recognition rates were given.

2

The Fricative recognition algorithms in Jassem (1979) were for Polish Fricatives. Only the recognition rates of the Polish Fricatives [f, v, s, z, ʃ, ʒ] have been quoted. The spectral shapes of these fricatives are similar to the corresponding English ones.

3

The recognition rates when the spectrum is divided into three partial areas, averaged across 3 speakers.

4

The recognition rates when spectrum divided into 4 partial areas, averaged across 3 speakers.

5

The recognition rates when spectrum divided 12 partial areas, averaged across 3 speakers.

6

The recognition rates when the spectral envelope is approximated by a 5th order polynomial, averaged across 3 speakers.

7

The recognition rate of McKinnon and Lee's Fricative recognition algorithm which used distance measures in $Z - Z'$ space.

8

The recognition rate of McKinnon and Lee's Fricative recognition algorithm which used distance measures in $Z - Z'$ space, AMAX and Z_B .

Table 7.5. The success rates of several fricative sorting algorithms.

TMSboard was used to digitized the speech (see sec 4.2). However unlike in the CASTT the speech was processed at 20 kHz. Several of the TMSboards were able to sample speech at 10 kHz and 20 kHz. There was a switch on the TMSboard used to select the sampling rate. These versions of TMSboards did not have an in-built low-pass filter but required an external low-pass filter to band-limit the speech. To digitize the speech, it was passed through a 9 kHz low-pass Kemo filter, quantized by the 12 bit A/D and then stored in the IBM-PC XT. The speech was recorded in an anechoic chamber.

7.6.0.6 Obtaining The Fricative Spectrum

The spectrum of the fricatives used in the recognition algorithm was the root mean square of the spectrum $S_{AVE}(f)$. This was calculated by

$$S_{AVE}(f) = \frac{1}{M} \sqrt{\sum_{m=0}^{M-1} S_m(f)^2} \quad (7.10)$$

where M was 19 and where $S_m(f)$ is the magnitude spectrum obtained by a 512 point Fast Fourier Transform (FFT) on speech samples windowed by a 512 sample hamming window. Each consecutive $S_m(f)$ was overlapped by 256 samples. $S_{AVE}(f)$ is the root mean square average of the magnitude spectrum over 0.256 seconds of a fricative signal. The resulting spectrum was then normalized to make the maximum amplitude in the spectrum one unit.

The spectra were calculated in SIGPROC, a signal processing package which runs on the departmental VAX.

Figure 7.4 gives typical examples of the spectra obtained from the database. Ignoring for the moment the spectrum below 1 kHz, the following is a descriptive account of the obtained spectra. The palatal fricatives [ʃ,ʒ] were all observed to have a major peak in the 3.4 kHz - 4 kHz region. The spectral envelope at 1 kHz was at a low magnitude. It remain low until the 2.3 kHz - 2.9 kHz region, where it rose steeply to the major peak, then dropped away gradually. By 9 kHz it had dropped away to virtually zero. The alveolar fricatives [s,z] were all observed to have a major peak in the 6.3 kHz - 8.3 kHz region. The spectral enveloped remained at a low magnitude till the 4.4 kHz - 5.7 kHz region; then it rose gradually to the major peak. The envelope then dropped off gradually. The magnitude of most alveolar spectra at 9 kHz was still significant.

The spectra of the labiodental and dental fricatives [f,v,θ,ð] were indistinguishable. The spectral envelope was virtually flat.

The overall spectral shapes were similar to those given in the literature, as can be seen by comparing figure 7.4 with figure 7.2. The positions of the major peaks above 1 kHz occurred within the frequency ranges expected for each of the fricative groups (see table 7.4). In addition the overall shape for each of the unvoiced/voiced pairs was similar, as expected. The unvoiced and voiced spectra could be differentiated because harmonic pulses in the lower frequency region were present in the voiced fricative spectra but not present in the unvoiced fricative spectra.

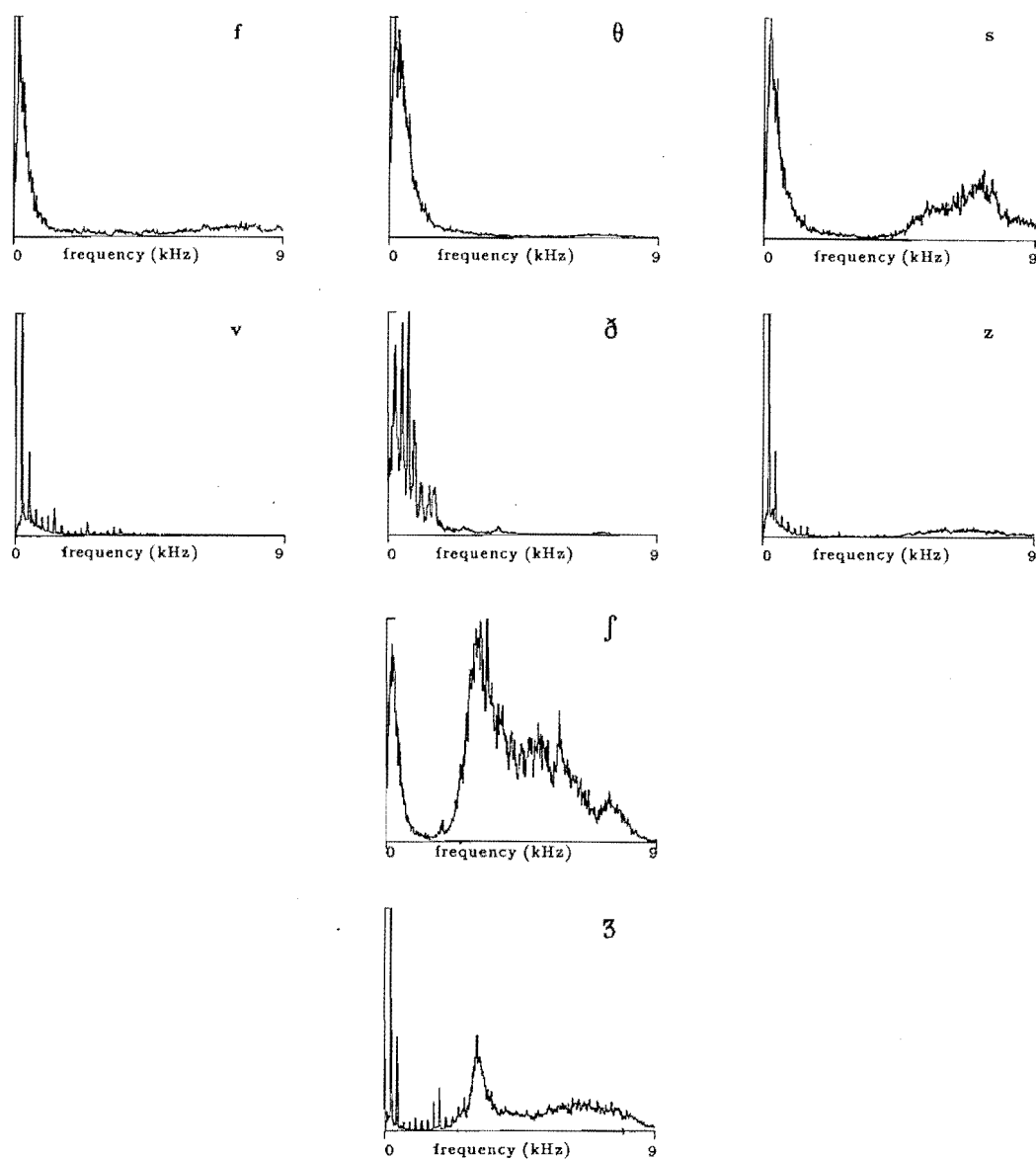


Figure 7.4. Typical spectral shapes of the eight English fricatives obtained from the isolated fricative.

7.6.1 The Algorithm

7.6.2 Labiodental/Dental, Alveolar and Palatal Distinction

Visual inspection of the spectra in fig 7.4 suggests it was possible to sort the fricative sounds into one of three groups, the palatals, the alveolars and the labiodentals/dentals. The position of the major peak above 1kHz in the spectrum is enough to identify the fricative type, see table 7.4. By dividing the spectrum into three bands and comparing the area under the spectral bands it was possible to roughly locate the position of the major peak and thereby identify the fricative type.

To achieve the sorting automatically the spectrum was partitioned into three frequency bands 0 to 2 kHz, 2 kHz to 6 kHz and 6 kHz to 9 kHz. If the maximum area was in the 2kHz to 6kHz band the fricative was identified as a palatal. If the maximum area was in the 6 kHz to 9 kHz band the fricative was identified as an alveolar. If the maximum frequency content was in the 0 kHz to 2kHz band the fricative was defined as either a labiodental or a dental.

The values of the frequency bands were obtained heuristically and were related to the positions of the major peaks. The major peak of the palatals was in the 2 kHz - 6 kHz band. The major peak in the alveolar spectrum was in the 6 kHz - 9 kHz band. Apart from the major peak at 200 Hz - 500 Hz the spectra of the labiodental/dentals fricative groups were virtually flat; thus the major peak for these was in the 0 - 2 kHz band.

7.6.2.1 Labiodental/Dental Distinction

As in the literature, (Baum and Blumstein, 1987) the spectral shapes of the labiodental and dental fricatives were found to be similar. For this reason they were identified as belonging to the same group in the first part of the algorithm. It has been suggested that the spectral information that distinguishes the two fricative types is in the 7kHz to 16kHz region (Harrington, 1988). Since the fricative sounds used in this investigation were sampled at 20 kHz and low pass filtered at 9 kHz, spectral information above 9 kHz was lost. To accentuate the remaining high frequency components in the 7 kHz to 9 kHz region the dental and labiodental fricative sounds were passed through a first order differentiator, that is the signals were pre-emphasised. This placed a 6 dB/octave lift on the spectrum.

Figure 7.5 shows typical examples of the spectra of labiodental and dental fricative obtained from the pre-emphasized signal. The spectra of the labiodental and dental groups are now visually distinguishable (c.f figure 7.4). The spectral envelope for the pre-emphasised dental fricative usually had two major peaks one in the 1 kHz region and the other in the 7 kHz - 8 kHz region and a valley within the 2 kHz - 6 kHz region. The spectrum of the pre-emphasised labiodental signal is approximately a straight line with a positive gradient.

To identify the fricatives from the spectral shape of the pre-emphasised signal the position in the spectrum of the major valley was found. The spectrum was divided into the frequency bands 0 kHz to 3 kHz, 3 kHz to 6 kHz and 6kHz to 9 kHz. If the minimum area was in the 3kHz to 6kHz frequency band the fricative was identified as a dental, due to the spectral valley observed in the 2 kHz to 6 kHz region of the pre-emphasized dental fricatives' spectrum. If the minimum area was in one of the

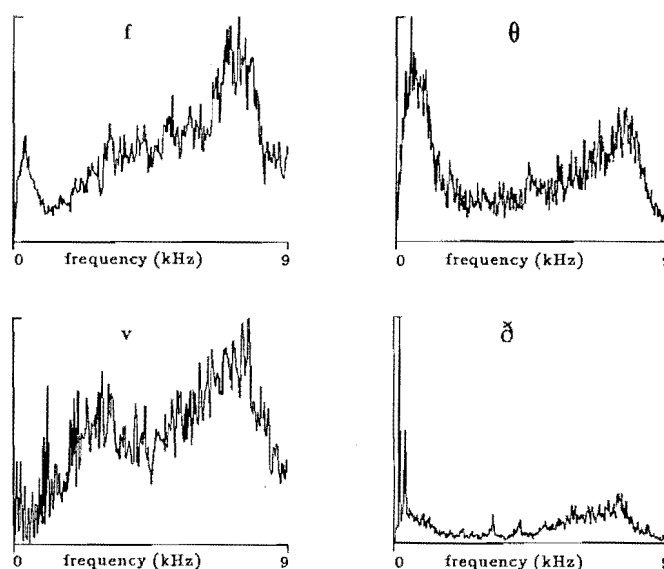


Figure 7.5. Typical examples of the spectra obtained from pre-emphasized utterances of [f,v,θ,ð].

other two frequency bands then the fricative was identified as a labiodental.

As for the Labiodental/Dental, Alveolar and Palatal distinctions, the frequency values for bands for the Labiodental/Dental distinctions were obtained heuristically.

7.6.2.2 The Voiced/Unvoiced Distinction

Once the fricative sound was classified according to place of articulation, as either a palatal, alveolar, labiodental or dental, it was necessary to decide whether it was voiced or unvoiced.

It was observed that in the spectra of the voiced fricatives there was a series of equally spaced spikes in the lower frequency regions (see figures 7.2 and 7.4). These spikes are the harmonics of the pitch. No similar spikes were observed in the unvoiced spectra due to the absence of periodic glottal excitation.

The pitch harmonics occur within the 0 - 2 kHz region. To distinguish between the voiced and unvoiced fricatives the spectra was band-limited to 2kHz and then a second FFT was taken. The area under the voiced fricative plots was much greater than that under the unvoiced fricative plots. If the area was greater than a preset threshold of 20 square units the fricative was identified as voiced, otherwise it was unvoiced. Thus the unvoiced and voiced fricatives could be distinguished.

Figure 7.6 gives the proposed fricative recognition algorithm. Table 7.6 gives the results of the recognition algorithm quoted in Brenda's final year report. The recognition rates vary from 72% to 100%. It can be seen that the recognition algorithm was very successful at recognising the fricatives [s,z,ʃ,ʒ], but it was not so successful at recognising [f,v,θ,ð]. However the recognition results were still very encouraging.

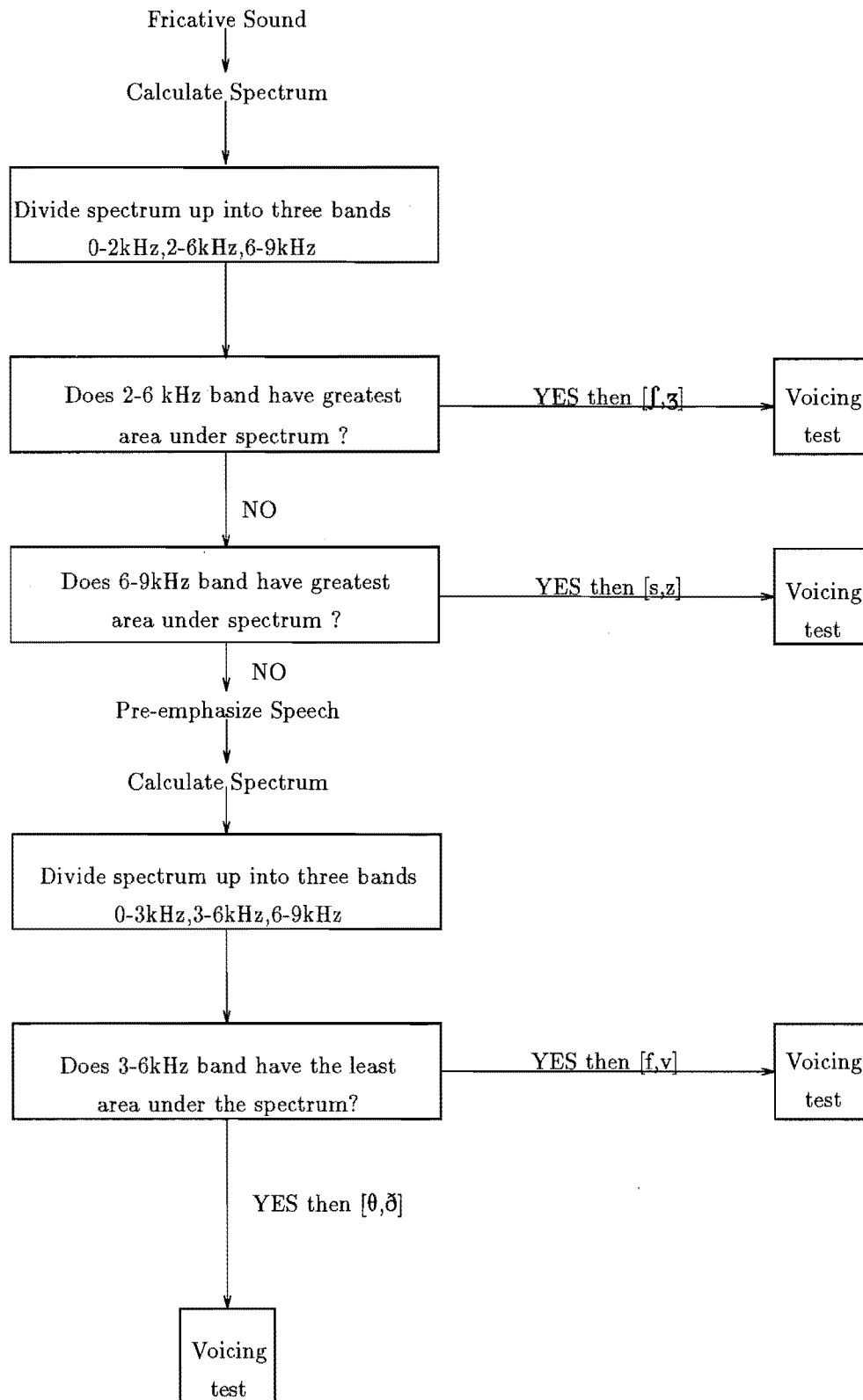


Figure 7.6. The Canterbury Fricative recognition algorithm for the eight English fricatives.

| fricative | f | v | θ | ð | s | z | ʃ | ʒ |
|------------------|------|------|------|-----|-------|-------|------|------|
| recognition rate | 85 % | 72 % | 85 % | 85% | 100 % | 100 % | 100% | 100% |

Table 7.6. The percentage of correct identification of fricatives using the proposed algorithm as reported in Satherley's final year report.

| | Initial | Medial | Final |
|---|---------|---------|--------|
| f | fat | coffee | reef |
| v | vat | river | dove |
| θ | thaw | ether | teeth |
| ð | that | rather | teethe |
| s | sue | recent | peace |
| z | zoo | visit | buzz |
| ʃ | shoe | fashion | push |
| ʒ | | leisure | rouge |

Table 7.7. The words from which the fricative portions were extracted for the study. Each word represents one of the fricatives in an initial, medial and final position. The fricative [ʒ] occurs only in medial and final positions in words in English.

7.6.3 Validating The Proposed Sorting Algorithm

To validate the proposed algorithm the author investigated the recognition rates obtained from a second fricative database. The recognition results of the algorithm with the new database will be presented and discussed below. First, however, the new database will be described.

7.6.3.1 Obtaining The Fricative Data

The fricative portions used to validate the algorithm were extracted from initial, medial and final positions in pre-recorded words, listed in table 7.7 from the Speech Group's speech database. At the time of the investigation the database consisted of pre-recorded speech of four New Zealand English speakers, two female and two male. For most of the words in the speech database two spoken versions per speaker existed. Thus for each fricative sound between ten and sixteen tokens were collected.

All the words in the database were recorded in the anechoic chamber to reduce the back ground noise. Speech was recorded using a standard microphone and digitized using the 16-bit analogue to digital converter on the SX10 board. All the speech in the database was sampled at 25 kHz. The cutoff point of the filter was 10.9 kHz.

7.6.3.2 The Results

The recognition results of the fricative recognition algorithm on the new set of data are given in tables 7.8. It can be seen that the algorithm correctly identified 90% of the labiodental and dental fricatives, 84% of the alveolar fricatives and 100% of the palatal fricatives. There are several reasons a recognition result is only given for the palatal, alveolar and labiodental/dental groups, rather than for the 8 individual fricatives for two reasons. Firstly the unvoiced/voiced decision did not work for fricative portions extracted from words. All the non-emphasised fricatives were classed as unvoiced and all the pre-emphasised fricatives were classed as voiced. Secondly the separation of the labiodental and dental fricatives was not at all successful. It can be seen from table 7.9 that only 54% of the labiodental fricatives were correctly identified and only 18% of the dental fricatives were correctly identified.

At first glance it could be construed that the recognition algorithm, when tested

| | | | |
|------------------|---------|-----|------|
| Fricative | f,v,θ,ð | s,z | ʃ,ʒ |
| Recognition Rate | 90% | 84% | 100% |

Table 7.8. The recognition results of fricatives portions extracted from the words listed in table 7.7 listed for the palatal, alveolar and labiodental/dental groups.

| | | |
|------------------|-----|-----|
| Fricative | f,v | θ,ð |
| Recognition Rate | 54% | 18% |

Table 7.9. The recognition results of fricatives portions extracted from the words listed in table 7.7 listed for the labiodental and dental groups.

on a new set of data, was not as promising as the results obtained with the isolated fricatives indicated it would be. However in hindsight it is unrealistic to expect the recognition rate for the fricatives which are extracted from words to be as good as the recognition rates for the isolated fricatives. This is for several reasons. It was difficult to ensure that only the fricative portion had been extracted from the words, especially for fricatives in medial positions. The extracted portions of the fricative were listened to, to ensure it was only the fricative portion of the word. However the short duration of some of the extracted signals made it difficult to ascertain if only the fricative portion of the word had been extracted. It was noted from visual inspection of the individual spectral shapes that the spectral shapes for the extracted fricative were not as distinctive as those resulting from the isolated sounds.

The second reason for the lower recognition rates is also related to the short duration of the fricative portions extracted from words. For voiced fricatives there is always a time delay before the glottis begins to vibrate, that is there is always an unvoiced portion. For voiced fricatives the proportion of the voiced portion to the unvoiced portion is much greater for fricatives uttered in sustained isolation than for fricatives in words. In some cases there is no voicing at all in the voiced fricatives extracted from words. Since the fricative portions were extracted from words, it was difficult to ensure that the voiced portion of the fricative had been extracted. Therefore it was not surprising that the algorithm could not distinguish between the unvoiced and voiced fricatives.

The third reason for lower recognition rates is due to the fact that many of the words with dental fricatives were actually pronounced with labiodental fricatives (e.g. the word "rather" was pronounced "raver"). This is not because the speakers who uttered the words suffered any speech impediments, but rather it is a phenomenon of New Zealand English. Unpublished studies by speech therapists have found the dental fricatives appear to be merging with the labiodental fricatives in New Zealand English¹. On relistening to the words in the speech group data base it was found of the 26 words that were intended to be pronounced with a dental fricative, 7 were pronounced with a labiodental fricative and for a further 8 it was not clear whether they were pronounced with a dental or labiodental fricative. Therefore about 60% of the words which were intended to be pronounced with dental fricative were not. This

¹Personal communication with Dr. M. Maclagan

| Recognition Method | f,v | s,z | ʃ,ʒ |
|-------------------------|--------|--------|--------|
| Hughes and Halle (1956) | 78.6 % | 79.8 % | 77.4 % |
| Jassem (1979) | 76 % | 93 % | 74 % |
| Canterbury (1993) | 94 %* | 92 % | 100 % |

Table 7.10. The recognition results of three fricative sorting algorithms, including the results from the Canterbury fricative sorting algorithm (* denotes that the dental fricative [θ,ð] are also included in this recognition result).

explains why the recognition rate for dental fricatives was so low (see table 7.9). It also explains why it was so difficult to distinguish between the dental and labiodental fricatives.

Therefore the recognition results obtained from the fricatives extracted from words can still be regarded as encouraging. In addition our recognition results compare very favourably with the results of the similar algorithms of Halle's and Hughes' (1956) and Jassem's (1979). Even when keeping in mind the discussion in sec 7.5.2.5 of the pitfalls of comparing results of recognition algorithms by different researchers and that Jassem's results are for the Polish fricative [f,v,s,z,ʃ,ʒ]. Our recognition results are given in table 7.10 along with those obtained by Hughes and Halle (1956) and Jassem (1979). To avoid confusion with the other fricative recognition algorithms, the algorithm developed by ourselves will be called the Canterbury fricative recognition algorithm.

7.7 THE IMPLICATIONS OF THE FRICATIVE RECOGNITION ALGORITHM FOR THE CASTT

The intention of investigating fricative recognition algorithms was, as stated at the beginning of the chapter, to find a new method of analysis for the newly proposed Fricative Monitor. The first conclusion that can be drawn from the investigation is if a new Fricative Monitor is to be built for the CASTT, the hardware of the CASTT itself will have to be updated. The bandwidth of the processed speech and therefore the sampling rate has to be increased. The highest frequency component in all the fricative spectra discussed in this chapter was at least 9 kHz. Currently in the CASTT it is 3 kHz.

Increasing the bandwidth of speech in the CASTT should not have a detrimental effect on the existing modules in the CASTT. The increased bandwidth will not affect the outputs of the Loudness Monitor or the Voice Pitch Tracker. It would change the outputs of the Spectrogram and Sustained Phonation modules. If a 0-9kHz spectrogram were plotted on the existing Spectrogram plots the output would look exceedingly coarse, since a frequency range of 9 kHz would be represented by only 64 points. However in the updated CASTT the graphics card would be at least an EGA card rather than a CGA card (the graphics card currently used in the CASTT). The increased screen resolution (640×320 c.f 320×200) means more detail can be plotted on the screen. The effects that increasing the bandwidth of speech would have on the Vocal Tract Shape module are not known. However the Vocal Tract Shape module will not remain in the CASTT in its current form, due to limitations in the existing analysis

algorithm. The proposed changes to this module will be discussed in the following chapter.

The results of the Canterbury fricative recognition algorithm were quite encouraging, especially for isolated fricatives. The fact that the recognition rates were much lower for fricatives extracted from words has no consequence for the newly proposed Fricative Monitor. The proposed Monitor is only intended as an articulation corrector for isolated fricatives. The VDT revealed that the Spectrogram module is quite an effective articulation corrector for sounds in words and sentences. The current 0-3 kHz bandwidth of the processed speech in the CASTT currently limits the spectrogram modules usefulness as an articulation corrector for fricatives (see sec 6.5.1.3). If the bandwidth of the speech is increased to 0-9 kHz it is quite likely that the Spectrogram module will have remedial potential for more of the fricative sounds. This would have to be tested, using the VDT, of course.

Therefore the newly proposed Fricative Monitor only has to recognise isolated fricatives. The Canterbury recognition algorithm developed by ourselves was quite successful in recognising the 8 English fricatives. However before the algorithm is incorporated into the new Fricative Monitor much more testing is required. Firstly the algorithm must be tested on a much larger data base than the 5 speakers it was originally tested on. There is currently no provision in the algorithm to reject sounds which are not fricatives; this needs to be included and tested. Finally the algorithm was only tested with isolated fricative sounds which were correctly pronounced. It was not tested to see whether it rejects or accepts common distortions of fricative sounds. For example if the algorithm identifies a lateral [s] as an [s], it is not selective enough. The algorithm must only recognise a sound, if the pronunciation of that sound falls within the limits of what is considered an acceptable pronunciation.

Whilst there is a large difference between the spectral shapes of the palatals, alveolars and labiodental/dentals, the difference in the spectral shapes between an acceptably pronounced fricative and unacceptably pronounced fricative is not likely to be very great. The place of articulation for the two is likely to be very similar (c.f. [s] with the lateral [s], both are alveolar fricatives - it is the tongue shape which is different). Pentz *et al.* (1979) said that the effective length of the vocal tract (the length from the constriction to the lips) had the greatest effect on the position of the major peak in the spectrum. Since the Canterbury fricative recognition algorithm recognises fricatives on the basis of the major peak it is quite possible that the Canterbury fricative recognition algorithm would not be able to distinguish between utterances in which the fricatives have been acceptably and unacceptably pronounced. If this were the case then several of the other methods for recognising fricatives, discussed in sec 7.5, could be investigated.

There is, therefore, a good deal of investigation still to be carried out before a new Fricative Monitor can be added to the CASTT.

CHAPTER 8

VOCAL TRACT SHAPE RECONSTRUCTION

The Vocal Tract Shape module was the second module of the CASTT in which both the VDT and the therapists' evaluations suggested the analysis algorithm required extensive changes (the first module being the Fricative Monitor). The VDT revealed that the Vocal Tract module could only reconstruct accurately the vocal tract shapes of [i] and [a] of both the female and male speaker. Several therapists in their evaluations felt the vocal tract reconstructions in the module were inaccurate, though they did not state explicitly whether this was for all vowel sounds or not.

This chapter presents the results of a preliminary investigation into a new method of reconstructing the vocal tract. However before that, the background of the vocal tract reconstruction problem will be discussed. A review of some of the existing vocal tract reconstruction algorithms will be given and the shortcomings of the current vocal tract reconstruction algorithm utilized in the Vocal Tract Shape module of the CASTT will be discussed.

8.1 A BRIEF INTRODUCTION TO SOME PAST RECONSTRUCTION TECHNIQUES

Reconstructing the vocal tract shape from the acoustic signal has long been a problem that has fascinated researchers (for example Mermelstein (1967), Schroeder (1967), Sondhi and Gopinath (1971), Wakita (1973), Sondhi and Resnick (1983), Lefevre *et al.* (1983), Milenkovic and Muller (1985)). Whilst it is a relatively straightforward task to construct the acoustic signal from a given vocal tract shape, the inverse problem is not so easy.

Many researchers have studied the vocal tract inverse problem but there are really only three distinct reconstruction methods that have been successfully employed. The first method is the impedance tube method. An impedance tube is placed at the lips and an impulse is sent down the tube into the vocal tract. Using the knowledge of the input and output signal the vocal tract shape is reconstructed. This method was used by Schroeder (1967), Sondhi *et al.* (1970),(1971),(1974),(1983),(1984) and by the group based predominantly at Laval University, Quebec (Descout *et al.*, 1976; Tousignant *et al.*, 1979; Lefevre *et al.*, 1980; Lefevre *et al.*, 1981; Lefevre *et al.*, 1983).

The second method is the direct estimation method in which the vocal tract is reconstructed from the acoustic speech signal measured by a microphone at the mouth. This method was first developed by Wakita (1973). Since then it has been widely used by subsequent researchers in vocal tract shape reconstruction. The final method reconstructs the vocal tract from measurements of the acoustic signal, via the microphone,

and measurements of the glottal pulse via an accelerometer at the throat. This method has been used by Vemula *et al* (1982) and Milenkovic (1984),(1985),(1987).

All three methods utilise the acoustic tube theory of the vocal tract. This theory will be outlined in the following section. The application of this theory in the vocal tract shape reconstruction algorithms of Sondhi *et al*, the Laval University group, Wakita and Milenkovic will be discussed.

8.2 THE ACOUSTIC TUBE MODEL OF THE VOCAL TRACT

8.2.1 Relating the Vocal Tract Shape to the Acoustic Signals

The vocal tract is the system of cavities above the larynx (Crystal, 1980, p94). A theory which relates the shape of the vocal tract to the acoustic signals is the acoustic tube model of the vocal tract. This model is the basis of all the vocal tract reconstruction algorithms. Therefore the acoustic tube model of the vocal tract will be presented before the different reconstruction algorithms are discussed.

In the acoustic tube model the vocal tract is modelled as a lossless, rigid walled, acoustic tube (losses due to wall vibration, viscosity and heat conduction are ignored). It is assumed the air in the tube behaves as a perfect gas and the wave propagation down the tube is linear and planar. Therefore the only sounds that the model accurately represents are vowels (which must be non-nasalized) and semi-vowels. Obstruents contain turbulent flow so they cannot be represented using this simplified model and nasal sounds are produced with a parallel acoustic tube, so they cannot be accurately modelled either.

With the above assumptions the wave propagation simplifies to a one dimensional model, with the x-axis along the axis of the vocal tract. The acoustic pressure, $p(x, t)$, and the fluid velocity, $v(x, t)$ in the vocal tract are then related to each other by:

$$\frac{\partial p(x, t)}{\partial x} = -\rho \frac{\partial v(x, t)}{\partial t} \quad (8.1)$$

and

$$k \frac{\partial p(x, t)}{\partial t} = -\frac{\partial v(x, t)}{\partial x} \quad (8.2)$$

where ρ is the density of air and k is the adiabatic compressibility of air (Morse, 1948). (8.1) and (8.2) are called the conservation equations of momentum and mass in one dimension.

Now (8.1) and (8.2) do not contain any reference to the vocal tract area. However, by considering the bulk flow rate of the air particles $u(x, t)$ rather than the particle velocity (8.1) and (8.2) can be rewritten with an area term. The bulk flow rate, $u(x, t)$ is defined as:

$$u(x, t) = a(x)v(x, t), \quad (8.3)$$

with $a(x)$ being the cross-sectional area of the vocal tract. Rearranging (8.3) and substituting it in (8.1) and (8.2) yields:

$$\frac{\partial p(x, t)}{\partial x} = -\frac{\rho}{a(x)} \frac{\partial u(x, t)}{\partial t} \quad (8.4)$$

and

$$k \frac{\partial p(x, t)}{\partial t} = -\frac{1}{a(x)} \frac{\partial u(x, t)}{\partial x} - u(x, t) \frac{\partial}{\partial x} \frac{1}{a(x)} \quad (8.5)$$

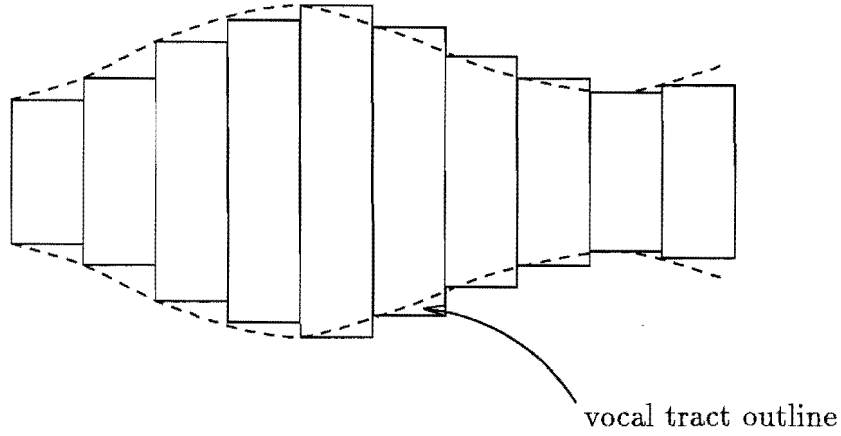


Figure 8.1. Approximating the vocal tract with a series of concatenated acoustic tubes.

The bulk flow version of the conservation of mass equation, (8.5), can be simplified by considering the nature of the cross-sectional area $a(x)$ to be recovered. The recovered cross-sectional area can be considered a continuous function, as it was in the case in Sondhi (1970),(1971),(1983). It is also assumed that the vocal tract shape does not change suddenly or rapidly (Mermelstein, 1967). Thus the rate of change of the cross-sectional area $a(x)$ in the x direction is very small, as is the rate of change for its inverse $\frac{1}{a(x)}$. With this in mind the second term on the right hand side of (8.5) can be considered much smaller than the first term and therefore can be neglected, yielding:

$$k \frac{\partial p(x, t)}{\partial t} = - \frac{1}{a(x)} \frac{\partial u(x, t)}{\partial x} \quad (8.6)$$

The alternative way to consider $a(x)$ is as a discrete function (Wakita, 1973; Lefevre *et al.*, 1983; Milenkovic and Muller, 1985). That is to approximate the vocal tract shape with a series of concatenated uniform tubes, as illustrated in fig 8.1. Thus for tube m

$$a_m(x) = a_m \quad (8.7)$$

In tube m (8.4) and (8.5) become:

$$\frac{\partial p_m(x, t)}{\partial x} = - \frac{\rho}{a_m} \frac{\partial u_m(x, t)}{\partial t} \quad (8.8)$$

and

$$k \frac{\partial p_m(x, t)}{\partial t} = - \left(\frac{1}{a_m} \frac{\partial u_m(x, t)}{\partial x} + u_m(x, t) \frac{\partial}{\partial x} \frac{1}{a_m} \right) \quad (8.9)$$

The second term in (8.9) is zero as a_m is a constant, therefore (8.9) becomes

$$\frac{\partial p_m(x, t)}{\partial x} = - \rho \frac{1}{a_m} \frac{\partial u_m(x, t)}{\partial t} \quad (8.10)$$

Using knowledge of the bulk flow and pressure waves measured it is possible to solve (8.4) and (8.6) or (8.8) and (8.10) to recover the shape of the vocal tract.

8.2.2 Relating The Reflection Coefficients Of The Concatenated Tube Model To The Vocal Tract Shape

Figure 8.1 shows the vocal tract represented as a series of concatenated uniform tubes. The conservation equations of momentum and mass, (8.8) and (8.10), hold within each tube and throughout the entire vocal tract. As a wave travels down the acoustic tube a certain amount of the wave will be reflected and a certain amount will be transmitted at each junction between adjacent cross-sectional areas, a_{m-1} and a_m say. The reflection coefficients at each junction are related to the two adjacent cross-sectional areas. For this derivation we will assume a_0 is at the lips.

Before the relationship between the reflection coefficients and the cross-sectional area is discussed, however, it is necessary to consider the nature of the pressure and bulk flow waves in the tube. This is done by finding the solution to the conservation equations.

8.2.2.1 Solutions To The Conservation Equations

In order to solve (8.8) and (8.10) it is easier to consider the pressure and bulk flow wave equations. By differentiating (8.8) with respect to x and (8.10) with respect to time and equating the like terms the pressure wave equation is found as:

$$\frac{\partial^2 p_m(x, t)}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 p_m(x, t)}{\partial t^2} \quad (8.11)$$

Likewise by differentiating (8.8) with respect to time and (8.10) with respect to x and equating the like terms the bulk flow equation is found as:

$$\frac{\partial^2 u_m(x, t)}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 u_m(x, t)}{\partial t^2} \quad (8.12)$$

where the quantity c , the speed of sound, is defined as:

$$c^2 = \frac{1}{k\rho} \quad (8.13)$$

Now the sinusoidal time dependent solution to (8.11) and (8.12) and hence (8.8) and (8.10) will consist of $Ae^{j\omega(t-x/c)}$ and $Be^{j\omega(t+x/c)}$ terms, where A and B are constants. Thus, from (8.11) we see the solution of the pressure is in the form

$$p_m(x, t) = p_m^+(t - x/c) + p_m^-(t + x/c) \quad (8.14)$$

and from (8.12) we see the solution of the bulk flow wave is of the form:

$$u_m(x, t) = u_m^+(t - x/c) - u_m^-(t + x/c) \quad (8.15)$$

where $p_m^+(t - x/c)$ and $u_m^+(t - x/c)$ are forward travelling components of the pressure and bulk flow waves and $p_m^-(t + x/c)$ and $u_m^-(t + x/c)$ are the backward travelling components.

In this thesis the positive x direction is the direction of the forward travelling wave components. Some vocal tract reconstruction methods have the forward travelling waves going from the lips to the glottis (for example Sondhi and Resnick (1983)), other methods have the forward travelling waves going from the glottis to the lips (for example Wakita (1973)). In the Direct Estimation Method employed by ourselves to

reconstruct the vocal tract and in the newly proposed vocal tract reconstruction method (see sec 8.4), the forward travelling waves are said to be going from the glottis to the lips. Therefore for this derivation relating the reflection coefficients of the concatenated tube model to the vocal tract shape (see sec 8.2.2) the forward travelling waves will also be going from the glottis to the lips.

8.2.2.2 Relating The Pressure and Bulk Flow Wave Components

Now the forward and backward travelling components of the pressure wave can be expressed in terms of the bulk flow forward and backward travelling waves. Using the relationship:

$$\frac{\partial f(t \pm x/c)}{\partial x} = \pm \frac{1}{c} \frac{\partial f(t \pm x/c)}{\partial t} \quad (8.16)$$

we find by substituting (8.14) and (8.15) into both (8.8) and (8.10) and then integrating with respect to time we can solve for $p_m^+(t - x/c)$ and $p_m^-(t + x/c)$ and find:

$$p_m^+(t - x/c) = \frac{\rho c}{a_m} u_m^+(t - x/c) + c_1 \quad (8.17)$$

$$p_m^-(t + x/c) = \frac{\rho c}{a_m} u_m^-(t + x/c) + c_2 \quad (8.18)$$

where c_1 and c_2 are constants of integration. They refer to the initial conditions, that is the static pressure. Since we are only interested in the pressure variations due to the travelling waves c_1 and c_2 can be ignored.

Being able to express the forward and backward components of the pressure wave in terms of the bulk flow means only one of the pressure and bulk flow signals needs to be measured in order to solve (8.8) and (8.10).

8.2.2.3 Behaviour Of Planar Waves At Boundaries Between Two Uniform Tubes

The pressure and bulk flow waves at each junction must be continuous. Thus at the junction between sections $m - 1$ and m :

$$p_m(x_{m-1}, t) = p_{m-1}(x_{m-1}, t) \quad (8.19)$$

$$u_m(x_{m-1}, t) = u_{m-1}(x_{m-1}, t) \quad (8.20)$$

where x_{m-1} is the distance between the junction and the lips.

Substituting (8.15) into (8.20) we get:

$$u_m^+(t - x_{m-1}/c) - u_m^-(t + x_{m-1}/c) = u_{m-1}^+(t - x_{m-1}/c) - u_{m-1}^-(t + x_{m-1}/c) \quad (8.21)$$

Substituting (8.17) and (8.18) into (8.14) and the result into (8.19) we get:

$$\frac{\rho c}{a_m} (u_m^+(t - x_{m-1}/c) + u_m^-(t + x_{m-1}/c)) = \frac{\rho c}{a_{m-1}} (u_{m-1}^+(t - x_{m-1}/c) + u_{m-1}^-(t + x_{m-1}/c)) \quad (8.22)$$

The equality expressed by (8.21) is illustrated in fig 8.2.

Consider the case of a single wave incident from the right (A_I in figure 8.3) (a). It will produce a transmitted wave, A_T , and a reflected wave, A_R . If $u_{m-1}^-(t + x_{m-1}/c)$ is set to zero then the forward wave $u_m^+(t - x_{m-1}/c)$, in fig 8.2 can be considered an

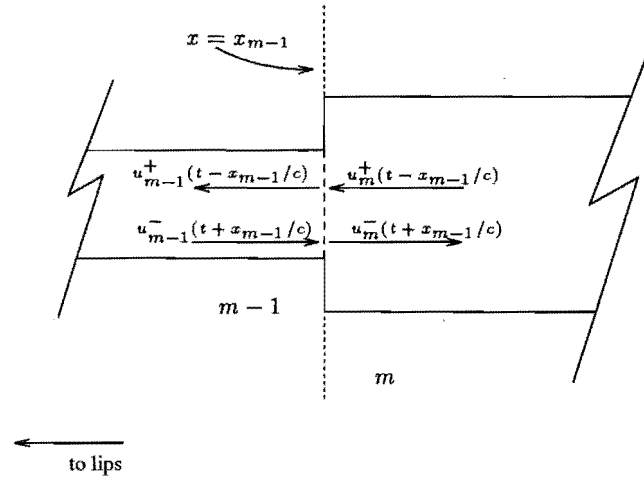


Figure 8.2. The behaviour of the bulk flow waves at the boundary between tube m and tube $m - 1$.

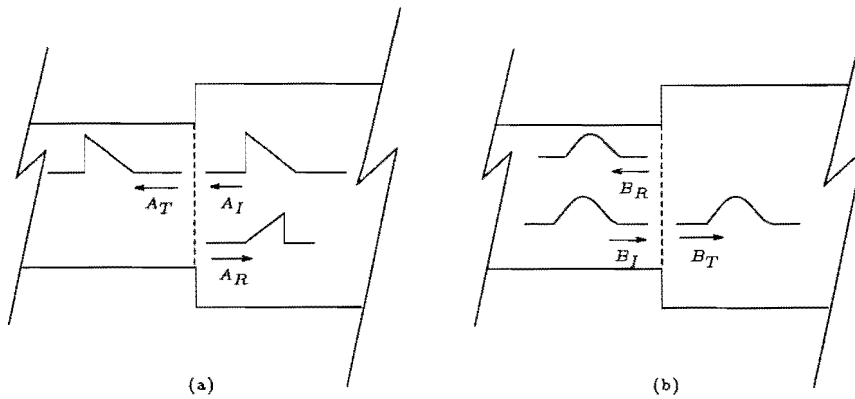


Figure 8.3. The behaviour of (a) the forward travelling wave A as it passes through a boundary between two adjacent tubes and (b) the forward travelling wave B as it passes through a boundary between two adjacent tubes.

incident wave, A_I , at the boundary $x = x_{m-1}$, the forward wave $u_{m-1}^+(t - x_{m-1}/c)$ the transmitted wave, A_T , and $u_m^-(t + x_{m-1}/c)$ the reflected wave, A_R . With $u_{m-1}^-(t + x_{m-1}/c) = 0$ (8.21) and (8.22) can be rewritten, giving

$$u_m^+(t - x_{m-1}/c) - u_m^-(t + x_{m-1}/c) = u_{m-1}^+(t - x_{m-1}/c) \quad (8.23)$$

and

$$\frac{\rho c}{a_m}(u_m^+(t - x_{m-1}/c) + u_m^-(t + x_{m-1}/c)) = \frac{\rho c}{a_{m-1}}u_{m-1}^+(t - x_{m-1}/c). \quad (8.24)$$

Eliminate the reflected wave $u_m^-(t + x_{m-1}/c)$ from (8.23) and (8.24), and rearrange to get:

$$u_{m-1}^+(t - x_{m-1}/c) = \frac{2a_{m-1}}{a_{m-1} + a_m}u_m^+(t - x_{m-1}/c) \quad (8.25)$$

Since $u_{m-1}^+(t - x_{m-1}/c)$ is the transmitted wave and $u_m^+(t - x_{m-1}/c)$ is the incident wave in (8.25), then

$$\frac{2a_{m-1}}{a_{m-1} + a_m} \quad (8.26)$$

is the transmission coefficient at the boundary $x = x_{m-1}$.

Now eliminate $u_{m-1}^+(t - x_{m-1}/c)$, the transmitted wave, from (8.23) and (8.24), and rearrange to get

$$-u_m^-(t + x_{m-1}/c) = \frac{a_{m-1} - a_m}{a_{m-1} + a_m}u_m^+(t - x_{m-1}/c) \quad (8.27)$$

where $u_m^+(t - x_{m-1}/c)$ is the incident wave and $-u_m^-(t + x_{m-1}/c)$ is the reflected wave.

$$\frac{a_{m-1} - a_m}{a_{m-1} + a_m} \quad (8.28)$$

is the reflection coefficient at the boundary $x = x_{m-1}$.

Similarly by setting $u_m^+(t - x_{m-1}/c) = 0$ the waves $u_{m-1}^-(t + x_{m-1}/c)$, $u_m^-(t + x_{m-1}/c)$ and $u_{m-1}^+(t - x_{m-1}/c)$ in figure 8.2 can be related to the incident wave B_I , transmitted wave B_T and reflected wave B_R in fig 8.3 (b), respectively. The transmission and reflection coefficients can be found for these waves by manipulating (8.21) and (8.22)

We will define μ_m to be the reflection coefficient which occurs between tube m and $m - 1$ as

$$\mu_m = \frac{a_{m-1} - a_m}{a_{m-1} + a_m} \quad (8.29)$$

Thus if the reflection coefficients, μ_m , $m = 1$ to M , M being the total number of acoustic tubes in the model, are found then it is possible to reconstruct the cross-sectional areas iteratively. The initial area must be known, otherwise it is only possible to recover the respective ratios between the cross-sectional areas.

8.2.2.4 A Gross Assumption Of The Concatenated Acoustic Tube Model Of The Vocal Tract

A gross assumption of this vocal tract model is the planar fluid flow. This is illustrated in figure 8.4. By representing the vocal tract as a series of discrete tubes a planar wave as it passes from tube α to β (see fig 8.4) is going to suffer disturbances in its flow as tube β is bigger than tube α . Thus in the model the length of the tubes should not be so long as to make big changes in the diameters between adjacent tubes.

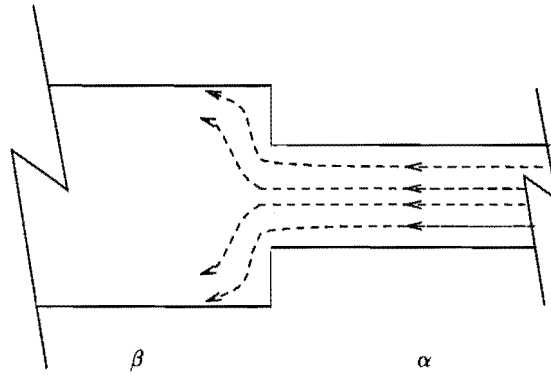


Figure 8.4. A fault with the acoustic tube model of the vocal tract.

8.3 VOCAL TRACT SHAPE RECONSTRUCTION METHODS

In order to solve (8.4) and (8.6) or (8.8) and (8.10) the initial conditions and boundary conditions must be set for the acoustic tube model. Each of the three reconstruction methods uses different criteria to set these conditions. These different conditions will be presented.

8.3.1 The Impedance Tube Method Of Vocal Tract Reconstruction

8.3.1.1 Sondhi's Method

Sondhi reconstructed the area of the vocal tract by exciting the vocal tract from an external source. Using the system illustrated in fig 8.5, the vocal tract was excited from the left end of the impedance tube with a series of acoustic pulses. There is no phonation, the glottis being kept closed. Measurements of the reflected and incoming acoustic signals were made at the right end of the tube, at the lips. This point is defined as $x = 0$. In this vocal tract reconstruction method the forward travelling waves go from the lips to the glottis. Therefore waves travelling from the lips to the glottis are travelling in the positive x direction (see fig 8.5).

Sondhi developed an algorithm for both the lossless and lossy vocal tract models. This section will only discuss the lossless algorithm; the reader is referred to Sondhi (1974) and Sondhi (1983) for a discussion on the lossy vocal tract model. In the lossless model the vocal tract is assumed to be a rigid wall tube, the air is a perfect gas and there are no losses due to wall vibration, heat conduction or viscosity. The wave motion is assumed to be linear and planar. Thus the continuity equations of momentum and mass, (8.4) and (8.6), hold. By choosing appropriate units for p and u , then the density ρ , the speed of sound c and the area at the lips, $a(0)$, can be set to unity in (8.4) and (8.6), simplifying them to:

$$\frac{\partial p(x, t)}{\partial x} = -\frac{\partial u(x, t)}{\partial t} \quad (8.30)$$

and

$$\frac{\partial p(x, t)}{\partial t} = -\frac{1}{a(x)} \frac{\partial u(x, t)}{\partial x} \quad (8.31)$$

These equations hold throughout the impedance tube and vocal tract. The cross-

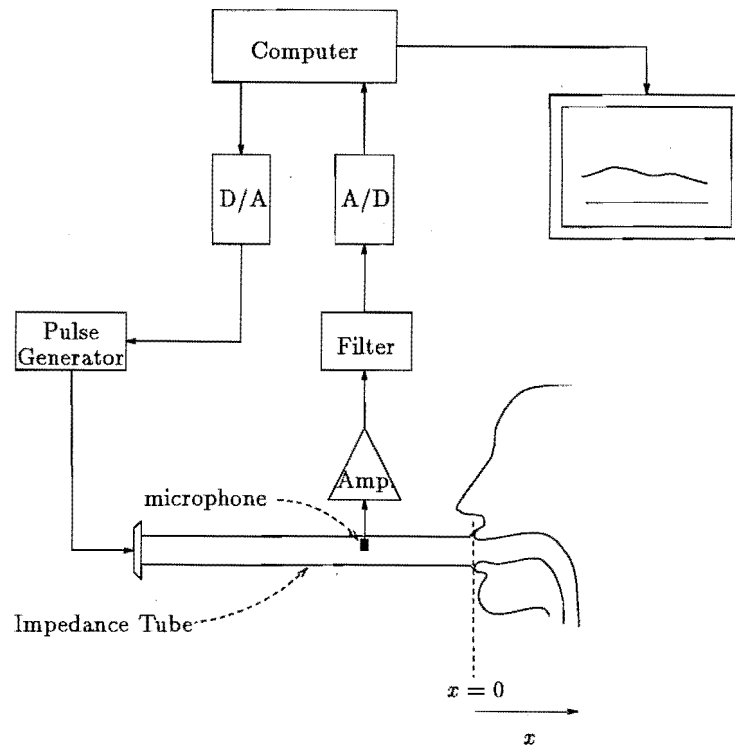


Figure 8.5. The system that Sondhi used for reconstruction of the vocal tract shape (Sondhi and Resnick, 1983; Sondhi, 1984).

sectional area of the impedance tube is:

$$a(x) = 1 \quad x \leq 0 \quad (8.32)$$

The origin, $x = 0$ is at the lips. It is assumed that the vocal tract is initially quiescent (Sondhi and Resnick, 1983). In addition it is assumed that the tract is stationary because “the motion is very slow on the time scale of acoustic phenomena of interest” (Sondhi, 1984).

Under these conditions Sondhi could reconstruct the vocal tract shape by calculating $a(x)$. He did it two ways; from the knowledge of the input impedance, $H(t)$, of the unknown cavity (the vocal tract) and from the Step reflectance, $S(t)$ of the vocal tract.

The input impedance is calculated from

$$p(0, t) = \int_0^t H(t - \tau) u(0, \tau) d\tau \quad (8.33)$$

where $p(0, t)$ and $u(0, t)$ are the pressure and bulk flow measured at the lips ($x = 0$). Now H is of the form

$$H(t) = \delta(t) + h(t) \quad (8.34)$$

where $\delta(t)$ is Dirac’s impulse function. Thus (8.33) becomes:

$$p(0, t) = u(0, t) + \int_0^t h(t - \tau) u(0, \tau) d\tau \quad (8.35)$$

By deconvolving the input impedance from (8.35) and using advanced mathematics, Sondhi was able to calculate the area $a(x)$ (Sondhi and Gopinath, 1971; Sondhi and Resnick, 1983; Sondhi, 1984).

Sondhi also derived the area from the step reflectance $\mathcal{S}(t)$. The step reflectance is defined as

$$\mathcal{S}(t) = \int_0^t \mathcal{R}(\tau) d\tau \quad (8.36)$$

where $\mathcal{R}(t)$, the reflectance, is defined as:

$$p_l(0, t) = \int_0^t \mathcal{R}(t - \tau) p_r(0, \tau) d\tau \quad (8.37)$$

here $p_l(0, t)$ and $p_r(0, t)$ are the left- and right-going waves at the lips, $x = 0$ respectively.

Sondhi (1983)(1984) summarizes the derivation of $a(x)$ from $s(t)$ as follows: For a quiescent tube at $t = 0$, let $u(x, t)$ and $p(x, t)$ be the solution of the nonlinear initial value problems:

$$\frac{\partial u(x, t)}{\partial x} + \frac{\partial u(x, t)}{\partial t} + \frac{u(x, x)}{p(x, x)} \left(\frac{\partial p(x, t)}{\partial t} + \frac{\partial p(x, t)}{\partial x} \right) = 0 \quad (8.38)$$

and

$$\frac{\partial u(x, t)}{\partial x} - \frac{\partial u(x, t)}{\partial t} - \frac{u(x, x)}{p(x, x)} \left(\frac{\partial p(x, t)}{\partial t} - \frac{\partial p(x, t)}{\partial x} \right) = 0 \quad (8.39)$$

in the region $x < t < 2L - x$ and $0 < x < L$, where L is the length of the vocal tract, with initial value at the lips of:

$$p(0, t) = 1 + \mathcal{S}(t) \quad (8.40)$$

and

$$u(0, t) = 1 - \mathcal{S}(t) \quad (8.41)$$

and the condition $u(x, t) = p(x, t) = 0$ for $t < x$. The area can then be calculated from the impedance:

$$a(x) = \frac{u(x, x)}{p(x, x)} \quad (8.42)$$

Sondhi was able to calculate the cross-sectional areas from the acoustic signal in 50ms. This was fast enough to display the dynamic variations of the tract in real time. Sondhi tested his vocal tract algorithm out on rubber moulds of vocal tract shapes and on two people. He concluded that for relatively open vocal tract shapes (for example [a]) the algorithm gave faithful reconstructions. But for vocal tract shapes with narrow constrictions, (for example [i]) the reconstruction of the area beyond the constriction (towards the glottis) was not correct (Sondhi and Resnick, 1983; Sondhi, 1984).

8.3.1.2 The Laval University Group Method

The group of researchers predominantly based at Laval University, (Descout *et al.*, 1976; Tousignant *et al.*, 1979; Lefevre *et al.*, 1980; Lefevre *et al.*, 1981; Lefevre *et al.*, 1983), have also developed a vocal tract reconstruction method using an impedance tube system similar to Sondhi's (see fig 8.5). An impedance tube is attached to the mouth. The glottis is kept shut and a pressure impulse is sent from the left end of the tube into the vocal tract. The pressure wave is measured as it goes into the vocal tract at the lips. The incident wave is denoted by $p_r(0, t)$ and $x = 0$ is at the lips. At some time later the resulting reflected pressure wave, $p_l(0, t)$, is measured at the same point. The time that elapses between the incident and reflected wave is small enough for the

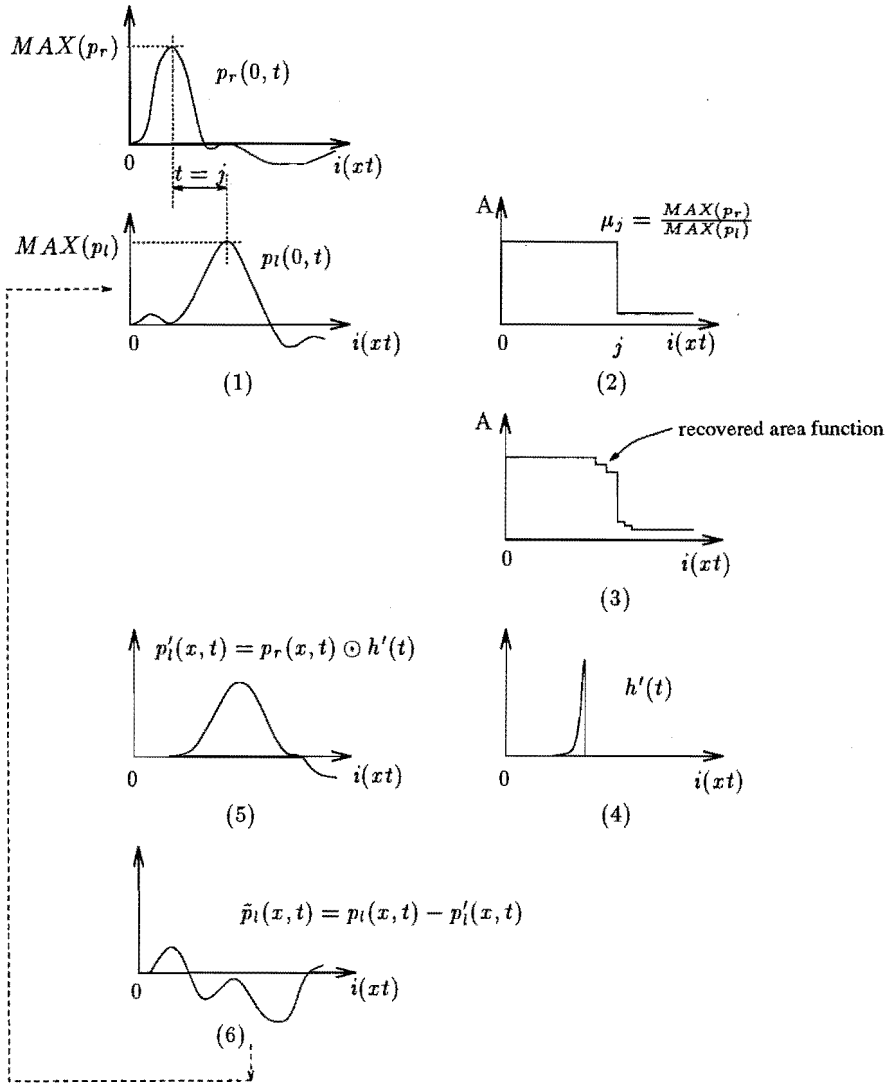


Figure 8.6. The vocal tract reconstruction method developed by the Laval University Group (from Lefevre *et al* (1983)).

vocal tract to be considered stationary. In this vocal tract reconstruction method, like Sondhi's (see sec 8.3.1.1), the forward travelling waves go from the lips to the glottis.

The Laval University group modelled the vocal tract by a series of lossless concatenated acoustic tubes. The length of each tube is much smaller than the wave-length of the highest transmitted frequency. They developed an iterative process to calculate the cross-sectional areas of the tubes. The method has been illustrated pictorially in fig 8.6. First the position of the reflection coefficient of the major constriction is found. This is done by calculating the time between the major peaks in the incident and reflected signal, see (1) in fig 8.6; call this time $t = j$. The position of the constriction is considered to occur approximately at $x = j$, due to the fact that the speed of sound is scaled to unity. The reflection coefficient μ_j is calculated from:

$$\mu_j = \frac{MAX(p_r)}{MAX(p_l)} \quad (8.43)$$

where $MAX(p_r)$ and $MAX(p_l)$ are the maximum amplitudes of the incident and reflected signals respectively (see (2) in fig 8.6). Since it is not physically possible for the

vocal tract to vary as abruptly as is illustrated in (2) fig 8.6, the edges of the reconstructed constriction are slightly softened. This is achieved by adding small steps both sides of the reconstructed constriction. The algorithm for doing this is described by Lefevre *et al* 1980. The resulting recovered vocal tract constriction is illustrated in (3) fig 8.6.

Next, using the transmission line model, the impulse response $h'(t)$ is obtained for the recovered area function (see (3) in fig 8.6). The output signal $p'_l(x, t)$ expected from the recovered area is calculated from the convolution:

$$p'_l(x, t) = p_r(x, t) \odot h'(t) \quad (8.44)$$

A residual output signal $\tilde{p}_l(x, t)$ is then calculated from:

$$\tilde{p}_l(x, t) = p_l(x, t) - p'_l(x, t) \quad (8.45)$$

Using the residual as the output signal the whole process is repeated again. In each iteration another constriction in the vocal tract is recovered. The method recovers the constrictions from the largest through to the smallest. The termination criterion of the iterative reconstruction process depends on the monitoring of four factors, the minimum mean square difference between the $p_l(x, t)$ and $\tilde{p}_l(x, t)$, whether values of the updated reflection coefficient values are beginning to diverge from their previous values, whether the area functions' values begin to oscillate and the sensitivity of the area function to the new iteration. If there are too many iterations, errors in calculating the area function will be introduced due to the noise in the recorded samples and the finite precision of the computations (Tousignant *et al.*, 1979). If there are too few iterations the vocal tract shape is not completely reconstructed.

The Laval University vocal tract reconstruction method was tested on real vocal tracts, as opposed to rubber moulds. From the existing knowledge of vocal tract shapes for vowel sounds the group found that the overall shapes reconstructed corresponded to the expected vocal tract shapes. The position of the constriction was well identified. However it was noted that the reconstruction of the area behind the major constriction was inaccurate, as observed by Sondhi (1983). The Laval Group noted also that some of the other errors in the reconstruction process could be overcome with practice. These errors were due to relaxing the muscular tension of the vocal tract, not placing the mouth in the correct position because there was no phonation, leakage at the lips because the impedance tube was not positioned correctly and bad control of glottal closure. They noted that the method was not good for nasal sounds or nasalized sounds. Nor was it good for sounds which involved narrowing the lips to less than 0.7 square centimetres or sounds which involved a lot of lip movement as the air-tight connection with the tube was lost. Sondhi made no mention of any problems he had using impedance tubes, but since he was mainly working with rubber moulds of vocal tracts rather than with people some may not have occurred.

8.3.2 The Direct Estimation Method

The direct estimation method of vocal tract reconstruction was first suggested by Wakita (1973). It has subsequently been used by many researchers in the vocal tract shape reconstruction problem (such as Brooks and Fallside (1976), Shigenaga *et al.* (1981), Aguilera *et al.* (1986), Guillemin (1986)). The method involves the linear

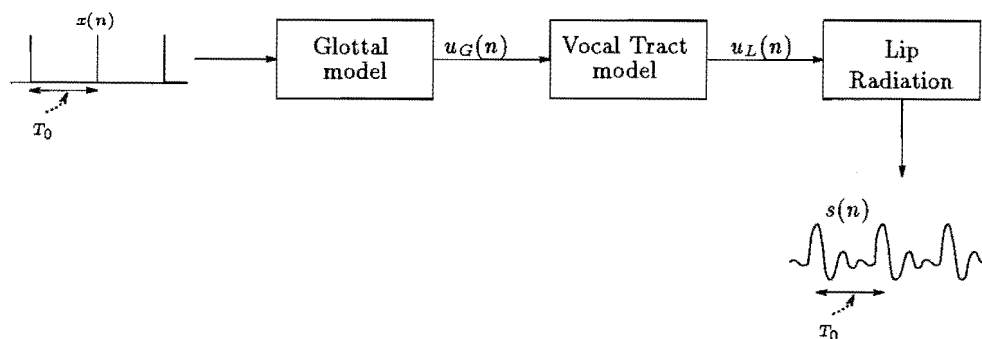


Figure 8.7. The linear model for speech production of non nasalized vowels and semi vowels.

model of speech production, the linear prediction model of speech and the acoustic tube model of the vocal tract. The last has been described in detail in section 8.2. The first two models will be discussed before the direct estimation by acoustic tube modelling of the vocal tract is presented.

In this vocal tract reconstruction method the forward travelling waves go from the glottis to the lips. Therefore waves travelling from the glottis to the lips are travelling in the positive x direction.

8.3.2.1 The Linear Model Of Speech Production

The linear speech production model enables one to model speech by considering it to result from a linear combination of the responses of the glottis, vocal tract and lips. Any interaction between the glottis and vocal tract, and vocal tract and lips is to be ignored. This model is illustrated in fig 8.7.

The vocal tract is modelled as a discrete time-varying all-pole filter, $V(z)$. It is assumed that the variations in the vocal tract occur slowly and that the slowly varying shape can be approximated by a succession of stationary shapes (Guillemin, 1986). The linear speech production model can account for both purely voiced non-nasal sounds and purely unvoiced speech. However due to the limitations of the acoustic tube model we are only interested in sounds in which the wave motion down the acoustic tube will be planar and linear, that is for non nasalized vowels and semi vowels. Therefore only the linear production model of speech for these sounds will be presented.

The behaviour of the glottis is described by

$$G(z) = \frac{1}{(1 - e^{-\omega_c T} z^{-1})^2} \quad (8.46)$$

ω_c is the cutoff frequency and T is the sampling period; it is customary to assume $T = 1$ (Markel and Gray, Jr., 1976). This glottal filter is excited by a unit impulse train, $X(z)$ with a period of T_0 , to produce a glottal waveform $u_G(n)$ (Guillemin, 1986). The frequency components of this waveform decay at a rate of -12 dB per octave (Witten, 1982). Next the signal $u_G(n)$ is input into the vocal tract filter $V(z)$, the output of which is $u_L(n)$. This waveform is then modified by the radiation at the lips. The radiation from the lips causes a 6dB lift per octave in the spectrum of the output signal (Witten, 1982) Thus the lips are modelled by $L(z)$, where:

$$L(z) = 1 - z^{-1} \quad (8.47)$$

and hence the linear speech production model for voiced speech $S(z)$ becomes:

$$S(z) = \Upsilon X(z)G(z)V(z)L(z) \quad (8.48)$$

where Υ is a gain factor which represents the variation in vocal intensity, $X(z)$ is an impulse train, $G(z)$ is a glottal shaping filter, $V(z)$ is the vocal tract filter and $L(z)$ models the effect of the lip radiation.

The vocal tract filter, $V(z)$, is described by the all-pole model:

$$V(z) = \frac{1}{\prod_{i=1}^K [1 - 2e^{-c_i T} \cos(b_i T) z^{-1} + e^{-2c_i T} z^{-2}]} \quad (8.49)$$

where K is the number of formants and where the i th formant frequency and 3 dB bandwidth are computed from $F_i = b_i/2\pi$ and $B_i = c_i/\pi$ respectively (Markel and Gray, Jr., 1976).

By combining (8.46), (8.47) and (8.49) and noting that the $e^{-\omega_c T}$ in (8.46) is approximately unity as the cut off frequency is 100 Hz (Guillemin, 1986), it can be seen that the term

$$\frac{1}{(1 - e^{-\omega_c T} z^{-1})}$$

in (8.46) virtually cancels the term

$$1 - z^{-1}$$

in (8.47). The result is a simplified model for $S(z)$:

$$S(z) = \Upsilon X(z) \frac{1}{A(z)} \quad (8.50)$$

where $A(z)$ is the inverse filter defined by:

$$A(z) = 1 - \sum_{i=1}^p \varphi_i z^{-i} \approx \frac{1}{G(z)L(z)V(z)} \quad (8.51)$$

here $p \geq 2K + 1$, and φ_i , $i = 1$ to p , are the filter coefficients (Markel and Gray, Jr., 1976).

8.3.2.2 The Linear Prediction Model of Speech

According to linear prediction analysis a speech sample $s(n)$ in the sampled time domain can be estimated from a weighted sum of past speech samples and some input function $x(n)$, that is:

$$s(n) = \Upsilon x(n) - \sum_{k=1}^p \varphi_k s(n-k) \quad (8.52)$$

where Υ is a gain factor which represents the variation in vocal intensity and p is the order of the model (Makhoul, 1975). In the direct estimation method the input $\Upsilon x(n)$ is unknown: thus $s(n)$ is approximated by:

$$\tilde{s}(n) = - \sum_{k=1}^p \varphi_k s(n-k) \quad (8.53)$$

where $\tilde{s}(n)$ is the approximation of $s(n)$. The φ_k coefficients, where $k = 1$ to p , in vector notation are represented as $[\varphi_p]$. For linear prediction to be a good model for speech it is desirable that the error, $e(n)$ between the real speech and its approximation:

$$\begin{aligned} e(n) &= s(n) - \tilde{s}(n) \\ &= s(n) + \sum_{k=1}^p \varphi_k s(n-k) \end{aligned} \quad (8.54)$$

is kept as small as possible. Thus it is necessary to find the set of $[\varphi_p]$, which minimizes the error $e(n)$.

Now taking the z-transform of (8.53) yields:

$$\tilde{S}(z) = F(z)S(z) \quad (8.55)$$

where $F(z)$ is:

$$F(z) = - \sum_{k=1}^p \varphi_k z^{-k} \quad (8.56)$$

which is the linear prediction filter (Markel and Gray, Jr., 1976) and $S(z)$ and $\tilde{S}(z)$ are real speech and approximated speech in the z-domain respectively. From (8.54) the linear prediction model in the z-domain is:

$$E(z) = S(z)(1 - F(z)) \quad (8.57)$$

Now comparing (8.50) with (8.57) we see that if $\Upsilon X(z)$ is set to equal $E(z)$ then $1 - F(z)$ is equal to $A(z)$ and hence the linear prediction model of speech results in an equivalent linear production model for speech (Markel and Gray, Jr., 1976).

The next step is to find the filter coefficients that minimize the error $e(n)$. The error is minimized using the least squares technique. The total squared error is given by:

$$\mathcal{E} = \sum_{n=n_0}^{n_1} e^2(n) \quad (8.58)$$

where n_0 to n_1 is the region over which the minimization will occur. The values of the two limits will be discussed later.

Substituting (8.54) into (8.58) and setting $\varphi_0 = 1$ gives:

$$\mathcal{E} = \sum_{n=n_0}^{n_1} \left(\sum_{k=0}^p \varphi_k s(n-k) \right)^2 \quad (8.59)$$

The set of coefficients $[\varphi_p]$, which now contains φ_0 , that minimizes \mathcal{E} is found by substituting (8.59) into (8.58), differentiating with respect to φ_i and setting the result equal to zero:

$$\frac{\partial \mathcal{E}}{\partial \varphi_i} = 2 \sum_{k=0}^p \varphi_k \sum_{n=n_0}^{n_1} s(n-i)s(n-k) = 0 \quad 1 \leq i \leq p \quad (8.60)$$

Now to find the solution of (8.60), which will give the $[\varphi_p]$ which minimizes \mathcal{E} , it is necessary to define the region over which the error is calculated, that is assign values to n_0 and n_1 . If \mathcal{E} is found over some finite region then the solution of (8.60) is found using the covariance method. This method will not be discussed because it was not used in the real-time vocal tract reconstruction algorithm. The interested reader is

referred to Markel and Gray (1976) for further discussion on this method. If the error \mathcal{E} is found over an infinite region the solution of (8.60) is found by the autocorrelation method. This is the method used in the real-time vocal tract reconstruction method (see sec 4.7.1), so it is the one that will be discussed. Whilst \mathcal{E} , in the autocorrelation method, is minimized over an infinite area, the speech outside a region 0 to $N - 1$ is set to zero by some window function; this sets the limits of the summation in (8.60) to $n_0 = 0$ and $n_1 = N - 1$.

By expanding out (8.60) it can be seen that

$$\frac{\partial \mathcal{E}}{\partial \varphi_i} = \sum_{n=0}^{N-1} s(n)s(n-i) + \sum_{k=1}^p \varphi_k \sum_{n=0}^{N-1} s(n-i)s(n-k) = 0 \quad 1 \leq i \leq p \quad (8.61)$$

By defining an autocorrelation function, $C_{|i-j|}$, as:

$$C_{|i-j|} = \sum_{n=0}^{N-1} s(n-i)s(n-j) \quad (8.62)$$

then (8.61) can be rewritten as

$$C_i + \sum_{k=1}^p \varphi_k C_{|i-k|} = 0 \quad 1 \leq i \leq p \quad (8.63)$$

$C_{|i-j|}$ is an even function, therefore $C_{|i-0|} = C_i$.

Now also by substituting (8.54) into (8.58), expanding, and then using (8.61), an expression for the error, \mathcal{E} is found:

$$\mathcal{E} = \sum_{n=0}^{N-1} s(n)^2 + \sum_{k=1}^p \varphi_k \sum_{n=0}^{N-1} s(n)s(n-k) \quad (8.64)$$

(this is, in fact, identical (8.59) Using the definition (8.62), (8.64) can then be rewritten:

$$\mathcal{E} = C_0 + \sum_{k=1}^p \varphi_k C_k \quad (8.65)$$

since $C_{|0-k|} = C_k$.

Now combining (8.63) and (8.65) we get the matrix relationship:

$$\begin{bmatrix} C_0 & C_1 & . & . & . & C_p \\ C_1 & C_0 & . & . & . & C_{p-1} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ C_p & C_{p-1} & . & . & . & C_0 \end{bmatrix} \begin{bmatrix} 1 \\ \varphi_1 \\ . \\ . \\ . \\ \varphi_p \end{bmatrix} = \begin{bmatrix} \mathcal{E} \\ 0 \\ . \\ . \\ . \\ 0 \end{bmatrix} \quad (8.66)$$

This autocorrelation matrix is a special form of matrix called a toeplitz matrix. Due to its symmetrical form about the diagonal, (8.66) has certain properties which enable it to be solved recursively. The recursive algorithm was devised by Durban and Levinson (Markel and Gray, Jr., 1976). The algorithm takes p iterations to solve (8.66) for $[\varphi_p]$, in each of these iterations a new estimate of \mathcal{E} being found. The $[\varphi_p]$ values will not be correctly solved for until the final step as they go through p updates. The

algorithm requires two initial conditions, firstly, in the first iteration the first estimate of the error set to:

$$\mathcal{E}_0 = C_0 \quad (8.67)$$

where the subscript on \mathcal{E} denotes the number of iterations the algorithm has gone through. The second initial condition is setting $\varphi_0 = 1$ for each iteration:

$$\varphi_{0,m} = 1 \quad (8.68)$$

where the subscript m denotes the m th iteration.

The Durban and Levinson algorithm iterates on the order m of the problem (Roberts and Mullis, 1987). To update each of the $[\varphi_{\mathbf{m}+1}]$ values in the m th iteration a quantity Γ_{m+1} is found

$$\Gamma_{m+1} = \frac{-\sum_{i=0}^m C_{m+1-i}\varphi_{i,m}}{\mathcal{E}_m} \quad (8.69)$$

Using this value the next estimate of the total error can be calculated:

$$\mathcal{E}_{m+1} = \mathcal{E}_m(1 - \Gamma_{m+1}^2) \quad (8.70)$$

Then the estimate of the filter coefficient $\varphi_{m+1,m}$ is set to:

$$\varphi_{m+1,m} = 0 \quad (8.71)$$

and the updated filter coefficients become

$$\varphi_{i,m+1} = \varphi_{i,m} + \Gamma_{m+1}\varphi_{m+1-i,m} \quad 0 \leq i \leq m+1 \quad (8.72)$$

Thus once the updated Γ_{m+1} is found from speech the filter coefficients for the next recursion level can be found. This process repeats for p iterations, at the end of which the filter coefficients which minimize \mathcal{E} are found.

8.3.2.3 Relating The Linear Prediction Model To The Acoustic Tube Model of Speech

After the m th iteration an estimate for the $m+1$ filter coefficients have been calculated, these $m+1$ values will be denoted by $[\varphi'_{\mathbf{m}+1}]$. Multiplying the $m+1$ $[\varphi'_{\mathbf{m}+1}]$ values by z^{-i} in (8.72) and then summing them up we get

$$\Omega_{m+1}^+(z) = \Omega_m^+(z) - \Gamma_{m+1}\Omega_m^-(z) \quad (8.73)$$

where $\Omega_m^+(z)$ is defined as

$$\Omega_m^+(z) = \sum_{i=0}^m \varphi_{i,m} z^{-i} \quad (8.74)$$

and $\Omega_m^-(z)$ is defined as

$$\Omega_m^-(z) = -\sum_{i=0}^m \varphi_{i,m} z^{-(m+1-i)} \quad (8.75)$$

Similarly by multiplying the $m+1$ $[\varphi'_{\mathbf{m}+1}]$ values by $z^{-(m+1-i)}$ in (8.72) and then summing them up we get

$$\Omega_{m+1}^-(z) = z^{-1}(\Omega_m^-(z) - \Gamma_{m+1}\Omega_m^+(z)) \quad (8.76)$$

where $\Omega_0^+(z) = 1$ and $\Omega_0^-(z) = -z^{-1}$.

The backward and forward going bulk flow components, $u_m^+(z)$ and $u_m^-(z)$, in the acoustic tube model are directly related to $\Omega_m^+(z)$ and $\Omega_m^-(z)$. In addition the predictor coefficients $[\Gamma_p]$ from the linear prediction model recursion algorithm are in fact equivalent to the reflection coefficients, $[\mu_M]$ in the acoustic tube model of speech. Using this knowledge and making several assumptions, it is then possible to reconstruct the vocal tract shape directly from the acoustic signal (Wakita, 1973), (Markel and Gray, Jr., 1976). In order to prove the equivalence between the set of Γ_i , $i = 0$ to p , coefficients from the linear prediction model and the set of $[\mu_M]$ reflection coefficients from the acoustic model it is necessary to relate the acoustic tube model of speech to the linear prediction model.

To do this it is necessary to add a time component into (8.21) and (8.22). In the direct estimation method the length of each of the concatenated acoustic tubes is constant and is of length l . Recall the boundary condition (8.20) of the bulk flow from figure 8.2. which can be rewritten:

$$u_m(l, t) = u_{m-1}(0, t) \quad (8.77)$$

By defining the time it takes the wave to travel from one end of a section of tube to the other, r , to be l/c , (8.21) and (8.22) can be rewritten:

$$u_m^+(t - r) - u_m^-(t + r) = u_{m-1}^+(t) - u_{m-1}^-(t) \quad (8.78)$$

and

$$\frac{\rho c}{a_m}(u_m^+(t - r) + u_m^-(t + r)) = \frac{\rho c}{a_{m-1}}(u_{m-1}^+(t) + u_{m-1}^-(t)) \quad (8.79)$$

Now by rearranging and manipulating (8.78) and (8.79) we are able to define both the forward and backward components of the bulk flow in tube m in terms of the forward and backward components of the bulk flow in tube $m - 1$ thus:

$$u_m^+(t - r) = \frac{1}{1 + \mu_m} (u_{m-1}^+(t) - \mu_m u_{m-1}^-(t)) \quad (8.80)$$

and

$$u_m^-(t + r) = \frac{1}{1 + \mu_m} (-\mu_m u_{m-1}^+(t) - u_{m-1}^-(t)) \quad (8.81)$$

where μ_m , the reflection coefficient between tubes m and $m - 1$, is defined by (8.29). By defining a time delay, t_m , which is the time a wave takes to flow from the right edge of tube m to the lips to be:

$$t_m = (m + 1)r, \quad (8.82)$$

then all the times can be given in reference to the lips. Then for the boundary between tube $m - 1$ and tube m , t is replaced with $t - t_{m-1}$. Since $t - t_{m-1} - r = t - t_m$ and $t - t_{m-1} + r = t - t_m + 2r$ then (8.80) and (8.81) can be rewritten:

$$u_m^+(t - t_m) = \frac{1}{1 + \mu_m} (u_{m-1}^+(t - t_{m-1}) - \mu_m u_{m-1}^-(t - t_{m-1})) \quad (8.83)$$

and

$$u_m^-(t - t_m + 2r) = \frac{1}{1 + \mu_m} (-\mu_m u_{m-1}^+(t - t_{m-1}) - u_{m-1}^-(t - t_{m-1})) \quad (8.84)$$

Now by defining the product:

$$\Theta_m = \prod_{i=1}^m (1 + \mu_i) \quad m > 0 \quad (8.85)$$

where Θ_0 is unity, we can calculate the reflection coefficient μ_m at the boundary between tube $m - 1$ and tube m recursively from the reflection coefficients $\mu_m, \mu_{m-1} \dots \mu_1$. By multiplying each side of (8.83) and (8.84) with (8.85) and remembering

$$\Theta_m = (1 + \mu_m)\Theta_{m-1},$$

we get

$$\Theta_m u_m^+(t - t_m) = \Theta_{m-1} \left(u_{m-1}^+(t - t_{m-1}) - \mu_m u_{m-1}^-(t - t_{m-1}) \right) \quad (8.86)$$

and

$$\Theta_m u_m^-(t + t_m + 2r) = \Theta_{m-1} \left(-\mu_m u_{m-1}^+(t - t_{m-1}) - u_{m-1}^-(t - t_{m-1}) \right) \quad (8.87)$$

Now if we define

$$\psi_m^+(t) = \Theta_m u_m^+(t - t_m) \quad (8.88)$$

and

$$\psi_m^-(t) = -\Theta_m u_m^-(t - t_m) \quad (8.89)$$

Then (8.83) and (8.84) can be rewritten as:

$$\psi_m^+(t) = \psi_{m-1}^+(t) + \mu_m \psi_{m-1}^-(t) \quad (8.90)$$

and

$$\psi_m^-(t + 2r) = \mu_m \psi_{m-1}^+(t) + \psi_{m-1}^-(t) \quad (8.91)$$

If (8.90) and (8.91) are sampled at $T = 2r$ then the Z-transforms can be obtained directly: the Z-transform (8.90) and (8.91) are:

$$\Psi_m^+(z) = \Psi_{m-1}^+(z) + \mu_m \Psi_{m-1}^-(z) \quad (8.92)$$

and

$$\Psi_m^-(z) = z^{-1}(\mu_m \Psi_{m-1}^+(z) + \Psi_{m-1}^-(z)) \quad (8.93)$$

Comparing (8.92) and (8.91) with (8.73) and (8.76) we see that:

$$\begin{aligned} \Psi_m^+(z) &= \Omega_{m+1}^+(z) \\ \Psi_m^-(z) &= \Omega_{m+1}^-(z) \end{aligned} \quad (8.94)$$

if M , the number of concatenated tubes in the acoustic tube model, is equal to p , the number of filter or predictor coefficients in the prediction model and if the boundary conditions of the acoustic tube model for the direct estimation method are equivalent to the initial and final conditions of the error recursion algorithm, then and only then

$$\mu_m = \Gamma_{m+1} \quad (8.95)$$

8.3.2.4 Defining The Boundary Condition On The Acoustic Tube Model When Using The Direct Estimation Method

In order to calculate the vocal tract cross-sectional shape by the the direct estimation method, several assumptions must be made in setting the boundary conditions. It is assumed that the behaviour of the acoustic signals at the termination of the acoustic tube model of the vocal tract, that is at the lips, can be described by saying the tube a_0 is connected to another section which has infinite area. This means the reflection coefficient at the lips is equal to one, $\mu_0 = 1$. Thus at the lips we get

$$u_0^+(t-r) = -u_0^-(t+r) \quad (8.96)$$

and the signal from the lips, $u_L(t)$, calculated using (8.15), is

$$u_L(t) = 2u_0^+(t-r) \quad (8.97)$$

It is also assumed that the acoustic tube is driven by a bulk flow source $u_G(t)$ whose source impedance is Z_G . By implication there is no contribution to the speech signal from any sublaryngeal components. In order to model the reflections at the glottis an artificial tube M , with area a_M , is added to the acoustic tube model (see figure 8.8). The glottis occurs at the right hand end of the tube $M-1$, see figure 8.8. The bulk flow source $u_G(t)$ is equal to the volume velocity going through the impedance Z_G plus that going into the acoustic tube

$$u_G(t) = \frac{1}{Z_G} p_{M-1}(0, t) + u_{M-1}(0, t). \quad (8.98)$$

Using (8.14), (8.15), (8.17) and (8.18) when $m = M-1$, the standard definition of the characteristic impedance for a tube m ($Z_m = \frac{\rho c}{a_m}$), equation (8.29) and the fact that the characteristic impedance at the glottis Z_G is defined as $\frac{\rho c}{a_M}$ we find:

$$u_G(t) = \frac{2}{1 + \mu_M} \left(u_{M-1}^+(t) - \mu_M u_{M-1}^-(t) \right) \quad (8.99)$$

By comparing (8.99) with (8.80) when m is equal to M it can be seen that

$$u_M^+(t-t_M) = u_G(t)/2 \quad (8.100)$$

Thus the volume velocity entering the glottis is $u_M^+(t-t_M)$.

In addition to the boundary conditions it is assumed that the effects of the glottal shape and lip radiation can be removed from the speech signal. Recalling (8.46) and (8.47) and that $e^{-w_c T}$ is approximately unity, it can be seen that $G(z)$ is approximately equal to $1/L^2(z)$. Thus (8.48) for $S(z)$ can be rewritten:

$$S(z) = \frac{V(z)X(z)}{L(z)} \quad (8.101)$$

Thus by pre-emphasising the speech signal $S(z)$ using

$$1 - z^{-1}$$

the resulting signal will be the vocal tract filter $V(z)$ times the quasi periodic pulse train $X(z)$. If it is assumed that by pre-emphasis the speech signal correctly accounts

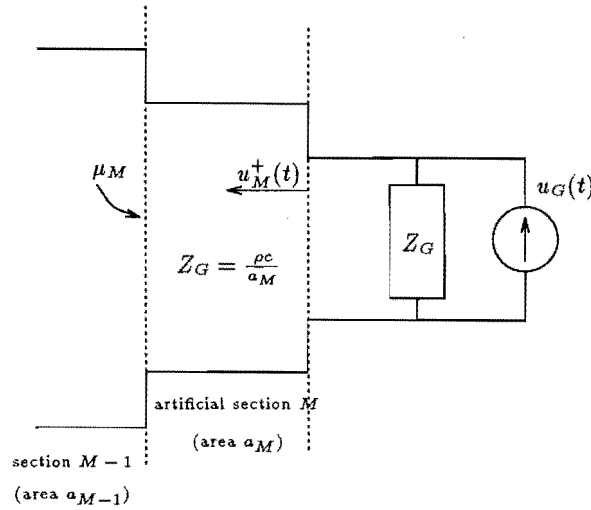


Figure 8.8. The boundary conditions at the glottis.

for the behaviour of the lips on the speech signal and removes the glottal pulse then it can be seen that

$$s'(n) = u_L(n) \quad (8.102)$$

where $s'(n)$ is the pre-emphasised speech and

$$e(n) = \frac{\Theta_M}{2} u_G(n - \frac{(M+1)}{2}) \quad (8.103)$$

and therefore we find

$$\mu_m = \Gamma_{m+1} \quad 1 < m < M$$

(Guillemin, 1986).

8.3.2.5 The Direct Estimation Vocal Tract Shape Reconstruction Method

The last several sections have discussed the parallels between the linear prediction model of speech and the acoustic tube model of speech. With this knowledge it is possible to reconstruct the vocal tract shape. The speech signal is first sampled and pre-emphasized using

$$1 - \nu z^{-1}$$

where ν is near unity. Next the set of filter coefficients $\varphi_i, i = 0$ to p which minimize (8.66) are found using the Durbin and Levinson recursion algorithm. In the process another set of coefficients $\Gamma_i, i = 1$ to $p+1$ are recovered. Since it has been proven that these coefficients are equivalent to the reflection coefficients in the acoustic tube model(see (8.95)), the cross-sectional areas for $m = p$ to 1 for the reconstructed vocal tract are recovered recursively using:

$$a_{m-1} = a_m \frac{\mu_m + 1}{\mu_m - 1} \quad (8.104)$$

where a_p , the glottal area, is unity. This is the method used in the Vocal Tract Shape module in the CASTT.

8.3.3 Vocal Tract Shape Reconstruction From Acoustic Measures At Two Points

Vemula *et al* (1982) and Milenkovic (1984),(1985) have investigated the reconstruction of the vocal tract shape using the signals measured at the lips, via a microphone, and at the larynx (throat) using an accelerometer.

Vemula *et al.* (1982) estimated the supraglottal pressure excitation signal using an auto-regressive moving average model of the throat wall to develop a transfer function which related the vibrations at the throat to the excitation signal. The vocal tract transfer function was then determined from the measured speech signal and the estimated supraglottal pressure signal. They calculated the vocal tract transfer function using the acoustic tube model of the tract. No information is given about how this was achieved or how they went on to calculate the cross-sectional areas of the concatenated acoustic tube model.

Vemula *et al.* (1982) compared the reconstruction of two different vocal tract shapes using their method and the direct estimation method. They concluded that their two-point method gave better results (Vemula *et al.*, 1982).

Milenkovic developed a different algorithm to calculate the vocal tract transfer function from the two measured signals (at the lips and throat). Using the two signals Milenkovic determined the low order poles and zeros of the real part of the driving point admittance Y_{de} . The Fourier transform of this is the vocal tract frequency response (Milenkovic and Muller, 1985). It is assumed that the vocal tract can be modelled by a series of concatenated acoustic tubes of equal length, d , and varying diameter. The reflection coefficients between adjacent tubes are calculated using the recursion algorithm given in sec 8.3.2.2 equations and the definition (8.95). The autocorrelation of the pre-emphasized signal C_m is replaced with the estimate of the driving point admittance, Y_{de} , of the vocal tract shape. The cross-sectional areas of the tube are then calculated from (8.104), as in the direct estimation method.

The zeros of the driving point impedance Z_{de} were found by calculating the coherence between the throat accelerometer and microphone signals. Distinct dips in the coherence function were found at the frequencies where the zeros were expected to occur (Milenkovic and Muller, 1985). The poles can be obtained from the spectrum Z_{de} . Once these poles and zeros are found the poles and zeros of the even part of the driving point admittance can be found, since the poles and zeros of the Z_{de} are the zeros and poles of Y_{de} .

Milenkovic's technique was limited to sustained vowels and was very sensitive to nasalization. However for vowels pronounced by three subjects, from which clean estimates of the poles and zeros were obtained, the area constructions obtained were consistent between speakers and with x-ray data (Milenkovic and Muller, 1985).

8.3.4 Discussion Of Vocal Tract Reconstruction Methods

Three different methods of vocal tract reconstruction have been reviewed. All three are based on the assumption that the vocal tract can be modelled as a lossless acoustic tube. The reconstructions were obtained from the measurement of one or two acoustic signals.

The intention of this review was to find an accurate and consistent vocal tract reconstruction algorithm for a speech module in the CASTT. There is currently a

reconstruction module in the CASTT but we want to update it since the vocal tract reconstructions are only accurate for some vowel sounds (see sec 6.5.2.1 and sec 8.3.4.1). In addition to accuracy and consistency, the manner in which the reconstructions are obtained is an important consideration when developing a reconstruction method for a speech therapy tool. Ideally it would be good if the reconstructions could be obtained when a person is phonating freely, and speaking as naturally as possible.

The impedance tube method used by Sondhi's group and the Laval University group was the most extensively researched reconstruction method. Both groups reported a high success rate in the accuracy and consistency of the reconstructions. However the reconstructions were obtained using an impedance tube attached to the mouth, the glottis being kept shut and there being no phonation. It is questionable whether for a person with a speech disability there would be any "correspondence [between] the articulations produced [using the impedance tube method] and the articulations produced in normal speech" (Milenkovic, 1984). It is also questionable whether this method would enable carry-over, that is correct pronunciation of the vowel in question when the person is talking freely and outside the therapy clinic. Carry-over is the end aim in any speech therapy process. For these reasons it was decided not to use the impedance tube method in the updated vocal tract reconstruction module.

The direct estimation method is appealing in its simplicity. The reconstructions are obtained from measuring the speech signal via a microphone. This is a fairly unobtrusive method of acquiring the data needed for the reconstruction process; it would not inhibit natural speech. Thus the application of the method to speech therapy aid has strong appeal. In addition to ourselves, several researchers (such as Crichton and Fallside (1974), Shigenaga *et al.* (1981), Aguilera *et al.* (1986), Abdelhamied *et al.* (1987)) have applied the method to these types of aids. However the method has been severely criticised by Sondhi (1979). Milenkovic (1984) and Vemula *et al.* (1982) have also questioned aspects of the method. All three say the assumptions made about the glottal behaviour in the direct estimation method are wrong. The arguments for and against this method warrant further discussion, which will be given in sec 8.3.4.1.

The two-point method investigated by Milenkovic and Vemula *et al.* appears to have some potential as a reconstruction module. However it has not been extensively tested, unlike the work done by Sondhi. The major advantage of this method over the direct method is that the actual glottal pulse is removed directly from the signal. The two point method would have applications in a speech therapy tool. The user would have to have an accelerometer attached to their throat to measure the glottal vibrations and they would have to speak into a microphone. The reconstructions would be obtained from natural speech. I investigated the two-point method. These findings will be outlined in sec 8.4.

8.3.4.1 The Problems Associated With The Direct Estimation Method

The major problem with the direct estimation method is the fact that the glottal pulse is unknown (Sondhi, 1979), (Vemula *et al.*, 1982), (Milenkovic, 1984). The method assumes that the effect of the glottal pulse is removed from the speech signal if the speech signal is pre-emphasized. This is because pre-emphasis lifts the speech spectrum by 6 dB/octave. It is assumed that the pre-emphasis of the signal coupled with the 6 dB/octave lift caused by the effect lip radiation has on the speech spectrum effectively

cancels the -12 dB/octave drop off caused by the glottal pulse on the speech spectrum. Sondhi (1979) does not believe that making the speech spectrum “flat” ensures that the effects of the source and radiation characteristics have been removed from the speech spectrum. The actual radiation and source characteristics have to be removed for the resulting spectrum to be the transfer function of the lossless vocal tract. Therefore, he says, it is not possible to reconstruct the vocal tract from just the speech signal; a second acoustic measurement is needed (Sondhi, 1984). Wakita, himself, in a later paper suggested the direct estimation method would be improved if the actual glottal pulse were deconvolved from the signal (Wakita, 1979).

Another criticism of the direct estimation method is that since the losses are not accounted for, the input impedance is not known, and the area functions are not unique (Sondhi, 1979). This problem does not arise in the impedance method because if the real part of the input impedance can be calculated for a lossless tract, the recovered area function is unique (Sondhi, 1979). Since the input signal is not known in the direct estimation method, the input impedance cannot be calculated.

No mention was made by Crichton and Fallside (1974), Shigenaga *et al.* (1981), Aguilera *et al.* (1986), Abdelhamied *et al.* (1987) about the accuracy and consistency of the vocal tract reconstruction displays in their aids. Thus no rebuttal to Sondhi’s criticisms of the reconstruction method can be drawn from their work.

Reconstructed vocal tract shapes from the Vocal Tract Shape module in the CASTT were scrutinized by a phonetics expert, Dr M. McLagan. Dr McLagan said the vocal tract reconstructions for the New Zealand front and central vowels [i,e,æ,ə] were good but the reconstructions for the New Zealand open and back vowels [o,ɔ,p,u,a] were not. The reconstructions were calculated from my voice. In all cases the reconstructions of the vocal tract shape from each vowel sound were repeatable. Whilst these findings indicate our reconstruction module is at least partially successful, they cannot strictly be considered as an endorsement of the direct estimation method. The reasons for this will now be outlined.

The reconstruction algorithm used in our vocal tract reconstruction module was given in sec 8.3.2.5. To briefly recapitulate, autocorrelation coefficients were calculated from the pre-emphasized speech signal. Then, using the Durban-Levinson recursion algorithm, (see sec 8.3.2.2), the reflection coefficients were calculated and from these the vocal tract shape was obtained. I discovered recently that the autocorrelation coefficients used in the Durban-Levinson recursion algorithm were being truncated severely in the TMS32010 digital signal processor (recall, from sec 4.7.1, that the Durban-Levinson algorithm was executed in the TMS32010 and the TMS32010 uses fixed point arithmetic). To overcome this anomaly, automatic scaling of the autocorrelation coefficients by a pre-calculated scale factor was incorporated into the reconstruction algorithm. This fixed the truncation problem, which meant we were now recovering the correct valued reflection coefficients. However the resulting reconstructions became completely inconsistent and inaccurate (before the algorithm had been partially successful).

Clearly the truncation of the eleven autocorrelation coefficients was affecting the accuracy and consistency of the reconstructions. Since the Durban-Levinson recursion algorithm is nonlinear, the effect that truncation of the autocorrelation coefficients has on the reflection coefficients will also be nonlinear. It is the lowest value autocorrelation coefficients which are corrupted the most due to the truncation; some of these will be set to zero. If the lowest value coefficients are the last ones to be calculated (e.g C_8, C_9

and C_{10}) then due to the recursive nature of the Durban-Levinson algorithm it will only be the last few reflection coefficients that will be affected by the corrupt autocorrelation coefficients. These last few reflection coefficients calculate the areas near the glottis in the vocal tract reconstruction. The occurrence of the low magnitude autocorrelation coefficients calculated from the speech signal will depend on the pitch of the speaker, the rate the signal is sampled at and the signal itself.

Even though we had partial success with our vocal tract reconstruction algorithm when the truncation of the eleven autocorrelation coefficients was occurring, that is, good reconstruction for New Zealand front and central vowels but inaccurate reconstructions for the back and open New Zealand vowels, the findings do not refute Sondhi's criticisms of the direct estimation method. The effect of truncation does not seem to be able to be linked to any acoustic phenomena. In addition it is difficult to see what direction one must take to improve the algorithm in order to increase the accuracy of reconstructing the back and open vowels. Thus with some reluctance it was decided not to pursue with the direct estimation method any further. Instead a new method of vocal tract reconstruction was investigated.

8.4 VOCAL TRACT RECONSTRUCTION BY INVERSE FILTERING - PRELIMINARY INVESTIGATION

The second method of vocal tract reconstruction investigated by myself was the two point acoustic measurement method. The method was first suggested by Richard Bates. The method investigated was slightly different from those developed by Vemula *et al* or Milenkovic. It was proposed that the glottal pulse can be removed from the speech signal by inverse filtering. To test this idea I used synthetic speech, $s'(t)$, which was obtained by convolving a known vocal tract filter, $v'(t)$, and a known excitation signal, $k'(t)$, thus

$$s'(t) = k'(t) \odot v'(t) \quad (8.105)$$

Now $k'(t)$ is defined as a known glottal pulse $g'(t)$ convolved with an impulse train $x(t)$.

The inverse filtering technique chosen was Wiener filtering. It is proposed that if the output of a suitable Wiener filter, $w_k(t)$, were convolved with the speech signal $s'(t)$ then the resulting signal would be the vocal tract filter $v'(t)$

$$v'(t) = s'(t) \odot w_k(t) \quad (8.106)$$

8.4.1 The Wiener Filter

The Wiener filter is a method of deconvolution, commonly used in image processing (Bates and McDonnell, 1986). The filter is made from the signal to be deconvolved, which in our application was the excitation signal $k'(t)$. The Wiener filter is defined as:

$$W_k(f) = \frac{K^*(f)}{|K(f)|^2 + \phi} \quad (8.107)$$

where $|K(f)|$ and $K^*(f)$ are the magnitude and complex conjugate, respectively, of the Fourier transform of $k(t)$, and ϕ is related to the signal-to-noise ratio. If $|K(f)|$ is much greater than ϕ then

$$W_k(f) \approx \frac{1}{|K(f)|}.$$

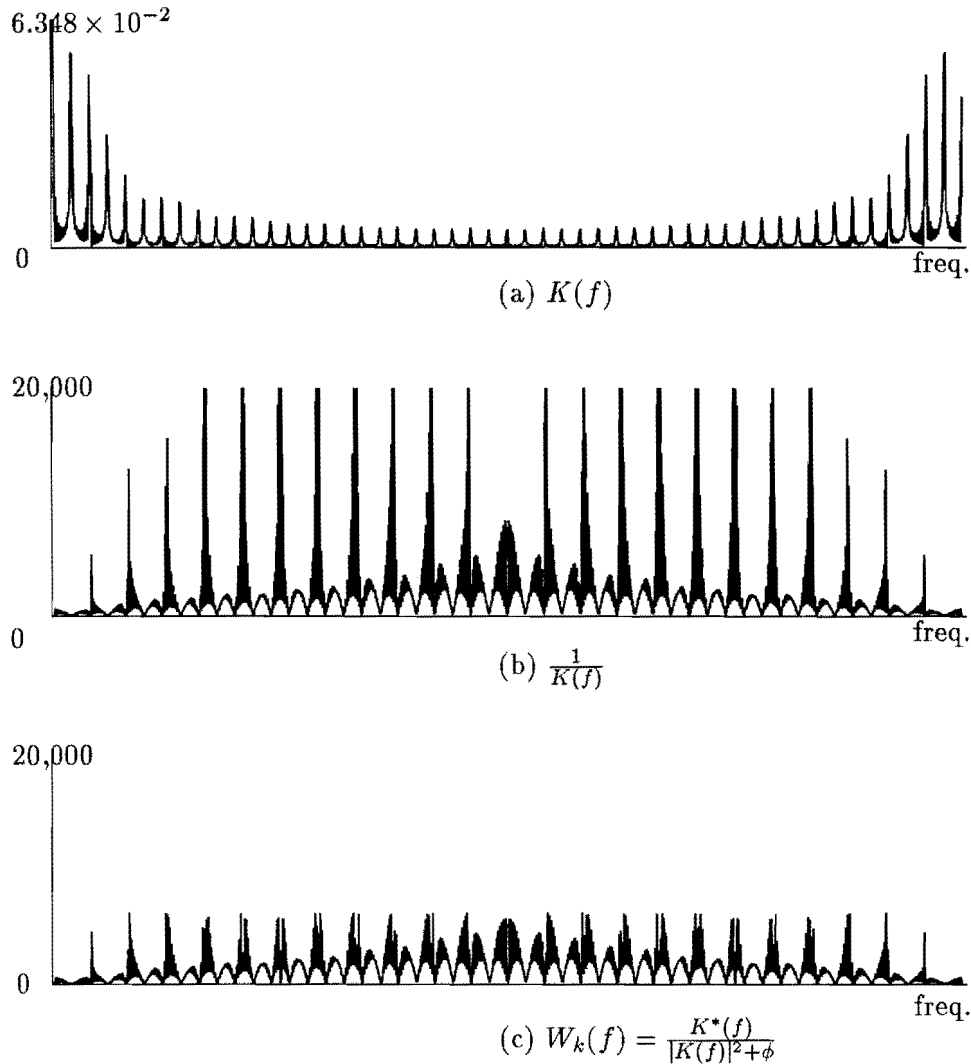


Figure 8.9. The effect of using a Wiener filter: (a) the original spectrum $K(f)$, (b) the inverse spectrum calculated using $\frac{1}{K(f)}$ (c) the inverse spectrum calculated using the Wiener filter $W_k(f) = \frac{K^*(f)}{|K(f)|^2 + \phi}$.

The presence of the ϕ component means that when the spectrum of $|K(f)|$ is exceedingly small in magnitude, it is still possible to calculate an inverse, without the inverse of the small magnitude components swamping the entire inverse spectrum. This can be seen in figure 8.9 by comparing the spectrum of $\frac{1}{K(f)}$ in fig 8.9(b) with the spectrum of the Wiener filter in fig 8.9(c). Figure 8.9(a) is the original spectrum $K(f)$.

$W_k(f)$ is transformed into the time domain, using the inverse Fourier transformation to obtain $w_k(t)$.

8.4.2 The Preparation For Vocal Tract Reconstruction By Inverse Filtering

As indicated in (8.105) the synthetic speech ($s'(t)$) was made by convolving one of the two vocal tract filters ($v'(t)$), filter 1 or filter 2, with one of the three different excitation signals ($k'(t)$), sawtooth 1, sawtooth 2 or sine-squared. For the rest of this chapter when the speech signal is mentioned, it will be the synthetic speech signal that

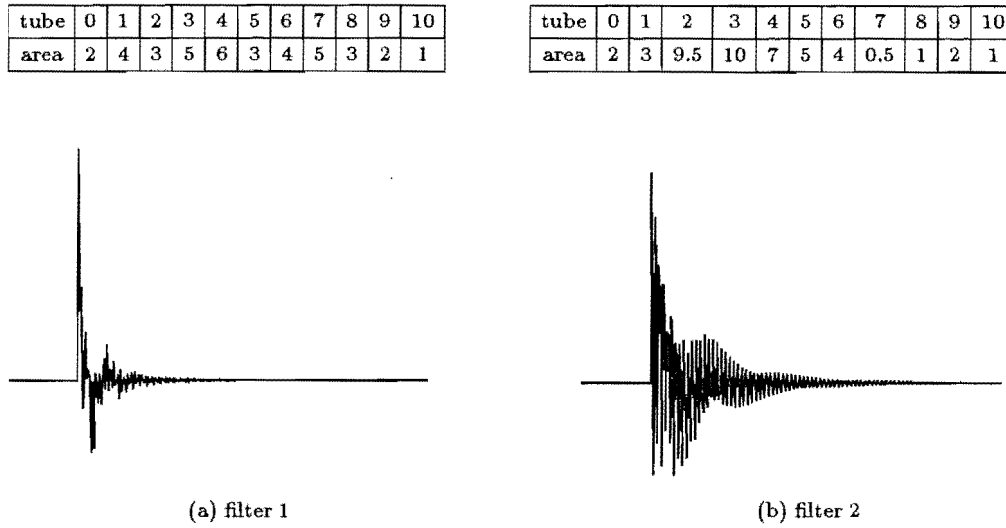


Figure 8.10. The cross-sectional areas of the vocal tract and the resulting filter response for (a) filter 1 and (b) filter 2.

is being referred to unless otherwise stated. The pitch of the excitation signals was either 100Hz or 200Hz, so 12 different (synthetic) speech signals were obtained. The vocal tract shapes and the corresponding filters, filter 1 and filter 2, are illustrated in fig 8.10. The vocal tract shapes were made up by the author. The excitation signals used are illustrated in fig 8.11: sawtooth 1 is a sawtooth pulse which lasted half the duration of the pitch period; sawtooth 2 is a sawtooth pulse which lasted a quarter the duration of the pitch period; and sine-squared is a sine-squared pulse which lasted half the duration of the pitch period. The sawtooth and sine-squared pulses were chosen because both are used as approximations of the glottal pulse (Witten, 1982).

A Wiener filter was made from the excitation signal. It was convolved with an extracted section of the speech signal. The number of pitch periods of both the speech signal and the excitation signal used in the Wiener filter must be the same. The resulting signal after the convolution process is the vocal tract filter response, see (8.106). The reflection coefficients are then recovered for this filter using the Durban-Levinson recursion algorithm, discussed in sec 8.3.2.2. In this case, however, the autocorrelation function C_i is calculated from $v'(t)$, rather than from the pre-emphasised speech signal. The areas of the concatenated acoustic tubes are recovered recursively using (8.104), where $m = 0$ to 9 and a_{10} is 1 (see fig 8.10).

8.4.2.1 The Guillemain Distance Measure

The accuracy of the recovered areas, compared to the true areas, is indicated by a distance measure. The distance measure used was the Guillemain distance measure, which is defined as:

$$d_g = \sqrt{\sum_{i=0}^m \left(\frac{a_i - a'_i}{1 + a_i + a'_i} \right)^2} \quad (8.108)$$

where a_i and a'_i are the areas of the i th tube of the original and reconstructed vocal tract shapes respectively. The distance measure was developed by Guillemain for the specific purpose of assessing the accuracy of reconstructed vocal tract shapes (Guillemain, 1986). The value of 1 in the denominator is a de-biasing factor. If the value of 1 were not

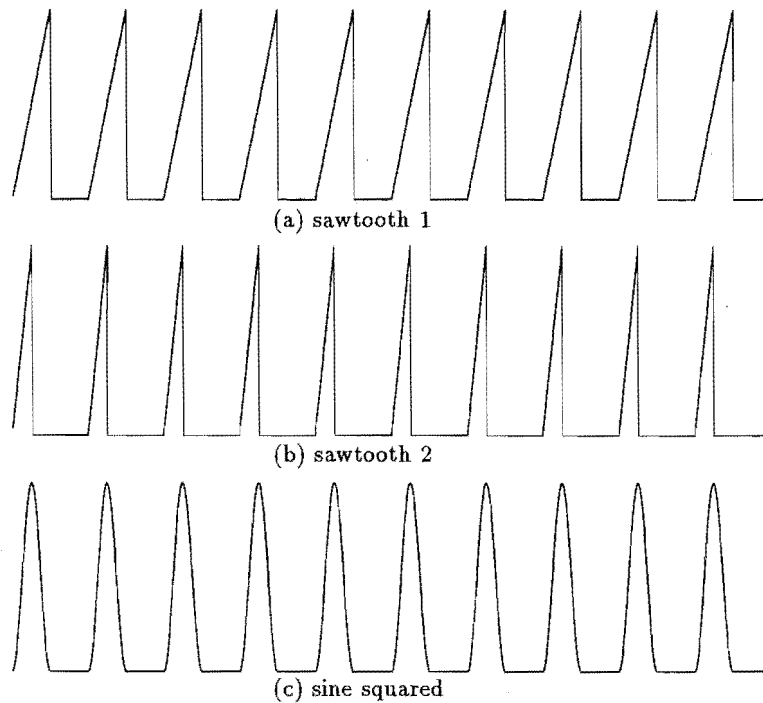


Figure 8.11. The excitation signals used in the experiment (a) sawtooth 1, sawtooth 2 and sine squared.

present then the measure would have a significant bias towards errors in small areas (Guillemin, 1986). The value of the de-biasing factor is arbitrary; Guillemin chose the value of 1 from experience. It was decided to use Guillemin's original de-biasing factor since both Guillemin and I modelled the vocal tract as a series of concatenated tubes and set the glottal area to 1 cm^2 .

The distance measure gave an indication of the goodness of the recovered shape. The smaller the distance measure the more accurate the reconstruction. From experience it was found that a d_g value of less than 0.1 indicates a very good reconstruction of the vocal tract shape and d_g values up to 0.5 still indicate good reconstructions of the vocal tract shape.

8.4.2.2 Making The Wiener Filter

There are two things that must be considered when making a Wiener filter: what the value of ϕ should be and the preparation of the signal from which the Wiener filter is to be made. To be able to recover the vocal tract filter response it is important to zero-extend the excitation signal in both directions before a Wiener filter is made. The importance of zero-extension will be discussed in the following section.

The next thing to consider in making the Wiener filter is the best value for ϕ . The effect ϕ has on the accuracy of the reconstructed vocal tract is illustrated in fig 8.12. It can be seen that as ϕ gets smaller so does the distance measure. Since we were only dealing with synthetic speech with no noise added, the only noise in the reconstruction process is going to be quantisation noise, which will be very small. For this reason ϕ

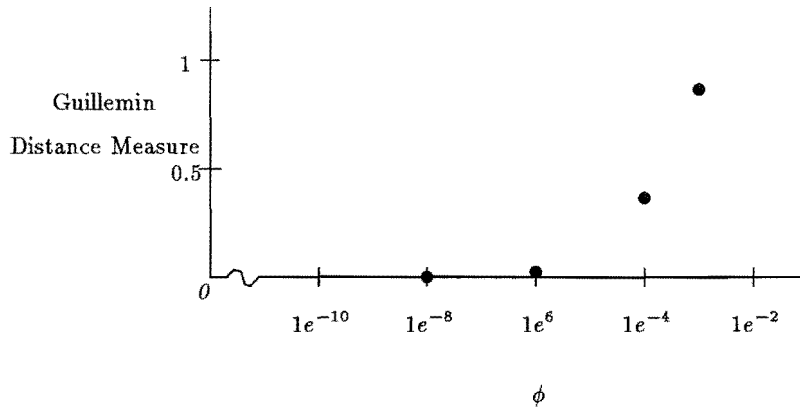


Figure 8.12. The effect ϕ has on the accuracy of the vocal tract reconstruction (in this case the signal was made from filter 1 convolved with sawtooth 1).

can be very small. ϕ was set to be 1×10^{-7} of the maximum amplitude of $|K(f)|$ for the sawtooth 1 and sawtooth 2 excitation signals and 1×10^{-11} of the maximum amplitude of $|K(f)|$ for the sine-squared excitation function.

8.4.2.3 The Importance Of Zero Extension

It is necessary to zero-extend all the signals in the convolution processes because the only convolution available in the signal processing package we used, was circular. All the signal processing in the vocal tract reconstruction investigation was performed in SIGPROC, a departmental signal processing package which ran on the departmental VAX. Circular convolution of two signals $x_1(n)$ and $x_2(n)$ is defined as

$$x_1(n) \odot x_2(n) = \sum_{k=0}^{N-1} x_1(n-k) \text{mod}_N x_2(k) \quad (8.109)$$

where N is the length of the convolved signal. For circular convolution to be equivalent to linear convolution the two input signals $x_1(n)$ and $x_2(n)$, of length N_1 and N_2 respectively, must be zero-extended until both are at least of length $N_1 + N_2 - 1$ (Ludeman, 1987). Zero-extending the time domain signals in order to do circular convolution does not add any more information to the signals (Ludeman, 1987), but it ensures that the output of the convolution does not wrap around on itself.

Since the vocal tract reconstruction was a deconvolution process, rather than convolution, the Wiener filter was not zero-extended but rather the excitation signal from which the Wiener filter was made was zero-extended. The speech signal, which was convolved with the Wiener filter, was also zero-extended. The importance of zero-extending the excitation signal is illustrated in fig 8.13. Figure 8.13(b) is just the extracted excitation signal, fig 8.13(e) is the excitation signal which has been zero-extended out to the right and in fig 8.13(h) the signal was zero-extended in both directions. A Wiener filter was made from each of the three excitation signals in fig 8.13 (b),(e) and (h). The Wiener filter outputs from these three signals are illustrated in the time domain in fig 8.13 (c),(f) and (i) respectively. The Wiener filter outputs in fig 8.13 (c),(f) and (i) are convolved with the extracted speech signal $s'(t)$ (see sec 8.13(a)) and the resulting vocal tract filter signals are seen in fig 8.13 (d),(g) and (j) respectively.

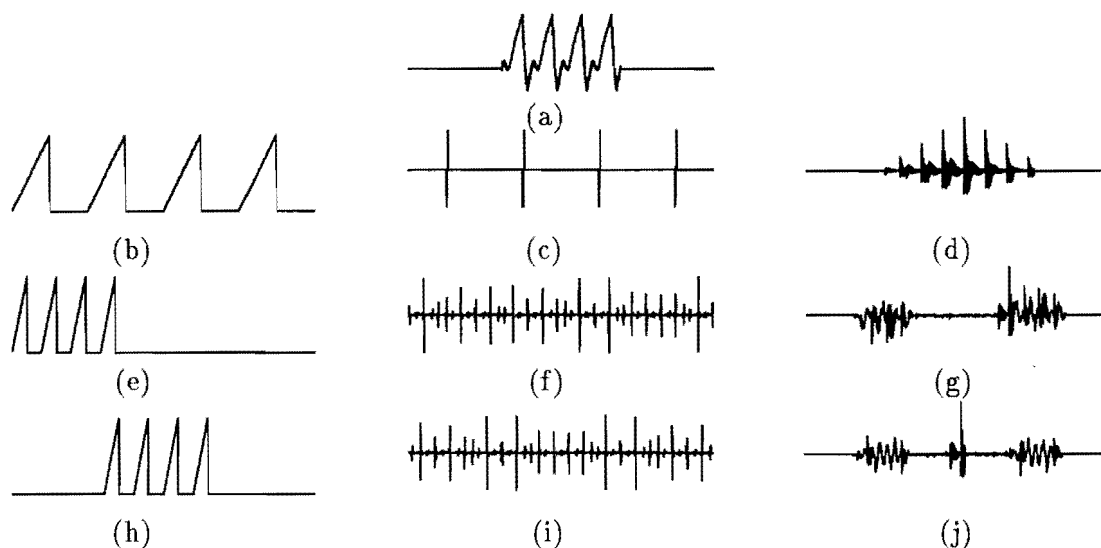


Figure 8.13. The effect that zero-extending the excitation signal to be used in the Wiener filter has on the reconstruction process for recovering the vocal tract filter response from a signal $s'(t)$ where (a) is the signal $s'(t)$, (b) is the excitation signal not zero-extended (e) is the excitation signal zero-extended to the right and (h) is the excitation signal zero-extended in both directions. (c), (f) and (i) give the resulting Wiener filters in the time domain of (b), (e). (h) and (d), (g) and (j) are the resulting filter response signals when the Wiener filters (c), (f) and (i) are convolved with the speech signal (a).

The convolved signal in fig 8.13(d) is not a good recovery of the vocal tract filter signal, because the filter output is lost within the quasi-periodic signal. It appears as a periodic, but truncated, version of the filter output of varying amplitude. This occurred because the major peaks of the both the signal and time domain Wiener filter, see fig 8.13 (b) and (c), are the same distance apart and reinforce each other in the convolution.

When the excitation signal is zero-extended the problem is removed. The major peaks of the Wiener filter in the time domain in fig 8.13 (g) and (j) are separated by a distance much greater than the pitch period. However the manner in which the excitation signal is zero-extended also must be considered. If the excitation is zero-extended in one direction only, as in fig 8.13(e), it can be seen in fig 8.13(g) that whilst the vocal tract filter has been recovered, it is partially submerged in the artifacts of convolution at the right of the convolved signal. The asymmetrical nature of the convolved signal is due to the asymmetric nature of the Wiener filter in the time domain, fig 8.13(f).

If the excitation signal is zero-extended in both directions as in fig 8.13(h), then the recovered vocal tract filter can be identified easily as it appears in the middle region of the convolved signal, see fig 8.13(j). To the left and right of the recovered filter are artifacts caused by the convolution process. Since it is a deconvolution process and we are only recovering a single filter response, we know our output signal will be shorter in duration than our two input signals. Before the next stage in the reconstruction process is performed, these artifacts must be removed by editing the signal.

8.4.3 The Effect Of The Excitation Function On The Accuracy Of The Vocal Tract Reconstruction

The reconstruction process was first investigated for six speech signals, which were made from each of the two filters convolved with each of the three excitation signals. All signals had a pitch period of 100 Hz. The resulting distance measures, indicating the accuracy of the reconstructions, can be seen in tables 8.1(a) and (b). The experiment established that it was possible to recover the vocal tract filter function from the speech signal by deconvolving the glottal pulse from the speech signal using signal processing. For signals which were made from either of the saw tooth excitation signals, the reconstructed vocal tract shapes were of much greater accuracy than the reconstruction from signals made from the sine-squared function.

In real life the actual shape of the glottal pulse is difficult to recover, as it is difficult to measure the waveform. The next experiment investigated the accuracy of the reconstructed areas if the Wiener filter was made from a different excitation signal than the speech signal. The resulting distance measures are also given in tables 8.1(a) and (b).

For speech signals made from the sawtooth 1 excitation signal and the Wiener filter made from the sawtooth 2 excitation signal or vice versa, the recovered areas were very close to the true areas, since the distance measures were less than 0.5 (see sec 8.4.2.1). If either the Wiener filter or the speech signal was made from the sine-squared excitation function then the recovered areas were not an accurate reconstruction: the d_g values for all instances were greater than one.

It was decided at this point not to continue using the sine-squared excitation signal as the accuracy of the reconstructed vocal tract was not as good as the saw tooth excitation signals. All the d_g values when the excitation, either of the signal or Wiener filter, was sine-squared was much greater than 0.5; hence the reconstructions were not very good.

8.4.4 The Effect of Zero-Extension On The Reconstruction Process

The extracted speech signal was zero-extended before it was convolved with the Wiener filter. In order to avoid discontinuities in the signal, it was zero-extended from a point where its actual amplitude was zero. There was usually more than one point in a period of the speech signals at which the amplitude was zero (see figure 8.14). It was found that the position of the zero point, relative to the major peak in the pitch period, from which the speech signal was extracted, affected the form of the recovered filter response. This in turn affected the accuracy of the recovered areas.

The effect the position of the point of zero amplitude had on the vocal tract filter form is illustrated in figure 8.14. Whilst the vocal tract filter is recognisable in all five of the reconstructions, they are not the same as the original. It is as if the beginning and the end of the response signal had been joined, to form a cycle and then broken at another position in the cycle, giving a new start and end point. It can be seen in figure 8.14 that as the points of zero amplitude move away to the left of the main peak in a speech signal pitch period, the "break" point in the hypothetical response signal cycle moves further away from the main spike in an anti-clockwise direction. The point of zero amplitude closest to the left of the major peak in the speech signal results in a recovered filter response where "cycle breakpoint" is almost to the immediate left of

| | excitation type in signal | | |
|----------------------------------|---------------------------|------------|--------------|
| excitation type in Wiener filter | sawtooth 1 | sawtooth 2 | sine-squared |
| sawtooth 1 | 0.132 | 0.180 | 1.39 |
| sawtooth 2 | 0.582 | 0.0956 | 2.01 |
| sine squared | 1.35 | 1.32 | 1.21 |

(a)

| | excitation type in signal | | |
|----------------------------------|---------------------------|------------|--------------|
| excitation type in Wiener filter | sawtooth 1 | sawtooth 2 | sine-squared |
| sawtooth 1 | 0.0732 | 0.344 | 1.84 |
| sawtooth 2 | 1.04 | 0.035 | 1.93 |
| sine squared | 1.64 | 1.71 | 1.79 |

(b)

Table 8.1. The effect of changing the excitation signal, used in the speech signal and in the Wiener filter, has on the accuracy of the vocal tract reconstructions, indicated by the Guillemín distance measure: (a) the distance measures for the speech signals obtained from filter 1, and (b) the distance measures for the speech signals obtained from filter 2.

| Position | d_g | Recovered Areas | | | | | | | | | | | |
|-----------|--------|-----------------|------|------|------|------|------|------|------|------|------|------|--|
| (a) | 0.132 | 1 | 2.25 | 3.37 | 5.71 | 4.39 | 3.34 | 6.76 | 5.30 | 3.21 | 4.23 | 2.09 | |
| (b) | 0.129 | 1 | 2.24 | 3.36 | 5.68 | 4.37 | 3.34 | 6.70 | 5.30 | 3.20 | 4.20 | 2.11 | |
| (c) | 0.130 | 1 | 2.24 | 3.38 | 5.67 | 4.37 | 3.35 | 6.67 | 5.27 | 3.25 | 4.15 | 2.13 | |
| (d) | 0.0891 | 1 | 2.23 | 3.26 | 5.36 | 4.22 | 3.27 | 6.21 | 4.94 | 3.15 | 3.88 | 2.14 | |
| (e) | 0.435 | 1 | 2.26 | 3.07 | 4.52 | 3.20 | 1.97 | 3.14 | 2.96 | 2.77 | 3.16 | 2.20 | |
| True area | - | 1 | 2 | 3 | 5 | 4 | 3 | 6 | 5 | 3 | 4 | 2 | |

Table 8.2. The distance measures resulting from the vocal tract reconstructions from the recovered vocal tract filter responses in 8.14.

the major spike. This recovered response looks the most similar to the original filter response, see fig 8.10 (a). In all five cases the same Wiener filter was used for the deconvolution.

The distance measures resulting from the vocal tract reconstructions from the recovered vocal tract filter responses are given in table 8.2, along with the original and recovered areas. The smallest distance measure is from the filter in position (d) ($d_g = 0.0891$). The largest distance measure is from the filter in position (e) ($d_g = 0.435$).

8.4.5 The Effect Of The Number Of Pitch Periods On The Area Reconstruction

The previous experiment was repeated on eight different speech signals. The speech signals were obtained by convolving either of the saw tooth excitation signals, which are at pitch frequencies of either 100 Hz or 200 Hz, with either of the vocal tract filter responses. For each of the eight signals two points of zero amplitude within one pitch

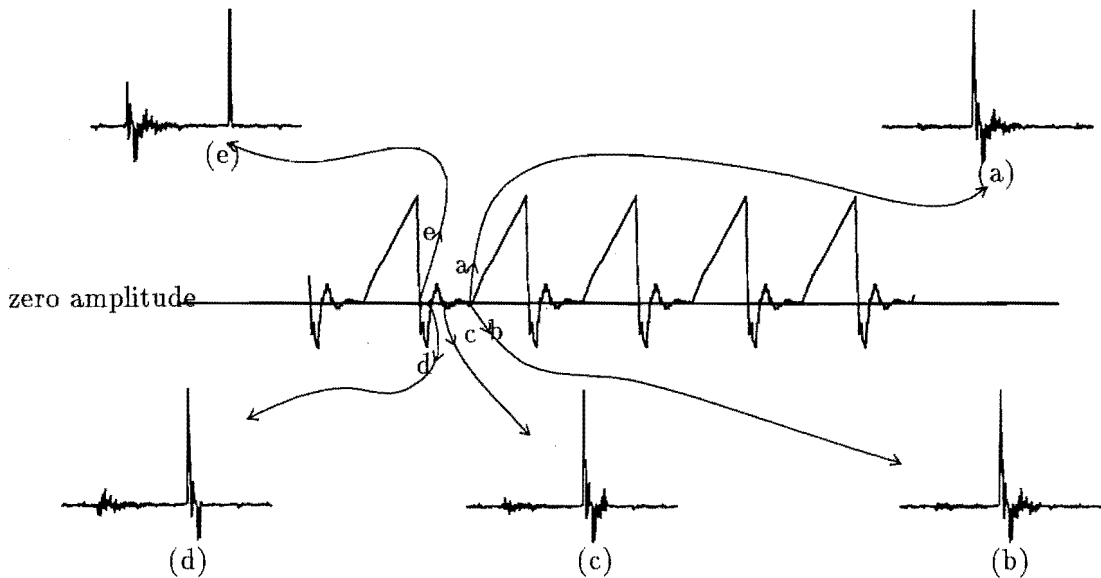


Figure 8.14. The effect that extracting the signal at different zero positions has on the form of the vocal tract filter response.

period were identified, the first zero point directly to the left of the major peak in a pitch period (e.g point (a) in figure 8.14), and the second zero point just to the right of the major peak in the pitch period (for example (e) in figure 8.14). From each of these points a portion of the speech signal was extracted and zero-extended, using the signal processing package SIGPROC. A Wiener filter was made from the same excitation signal which was used in the speech signal. The excitation signal had the same number of pitch periods as the extracted speech signal. The Wiener filter was convolved with the speech signal to get the vocal tract filter response. This process was repeated at each point of zero amplitude four times. From each point, distance measures were obtained for the areas reconstructed from four, three, two and one pitch periods of the speech signal.

The distance measures for the recovered area functions for the four 100 Hz speech signals are plotted in fig 8.15 (a), (b), (c) and (d). They are plotted for the four 200 Hz speech signals in fig 8.16 (a), (b), (c) and (d). The distance measures for each of the eight speech signals were given for the four, three, two or one pitch periods of the signal extracted from the different points of zero amplitude. The points marked with 'x' are the distance measures calculated from the vocal tract reconstructions when the signal is extracted at the zero point just to the right of the main peak. The points marked with '+' are the values of the distance measures of the signal extracted from the zero position just to the left of the main peak.

Several observations can be made about the reconstruction process from fig 8.15 and fig 8.16. In all cases it can be seen that the least accurate vocal tract reconstruction is obtained (largest d_g value) when the speech signal is extracted from the point of zero amplitude closest to, but to the right of the major peak within a pitch period of the speech signal, (such as point (e) in fig 8.14). The starting position from which the speech segment is extracted has a bigger effect on the accuracy of the reconstruction process than the number of pitch periods the reconstruction is performed over.

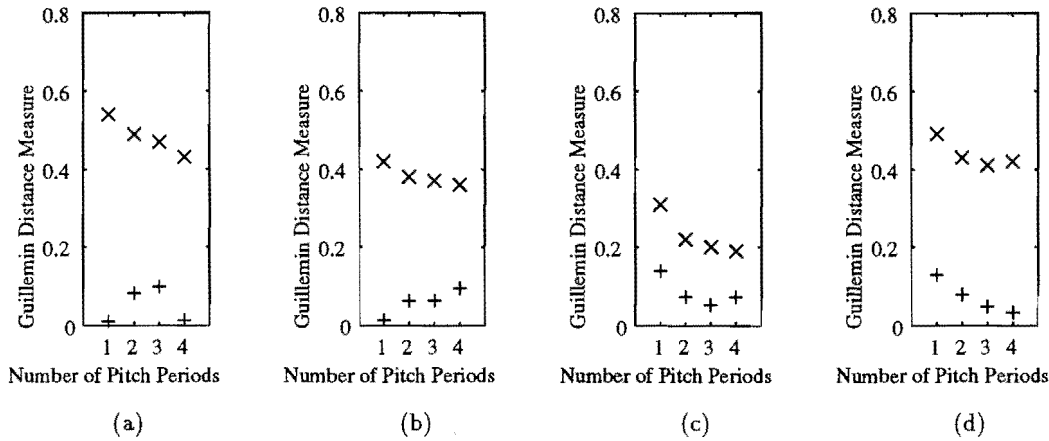


Figure 8.15. The distance measures resulting from the vocal tract reconstruction done from the four 100 Hz speech signals extracted from the original speech signal at the first zero position to the right of the main peak in the pitch period (indicated by 'x') and at the last zero position to the left of the main peak in the pitch period (indicated by '+') over four, three, two and one pitch periods of the signal for: (a) a 100 Hz signal made from filter 1 convolved with sawtooth 1, (b) a 100 Hz signal made from filter 1 convolved with sawtooth 2, (c) a 100 Hz signal made from filter 2 convolved with sawtooth 1, and finally (d) a 100 Hz signal made from filter 2 convolved with sawtooth 2.

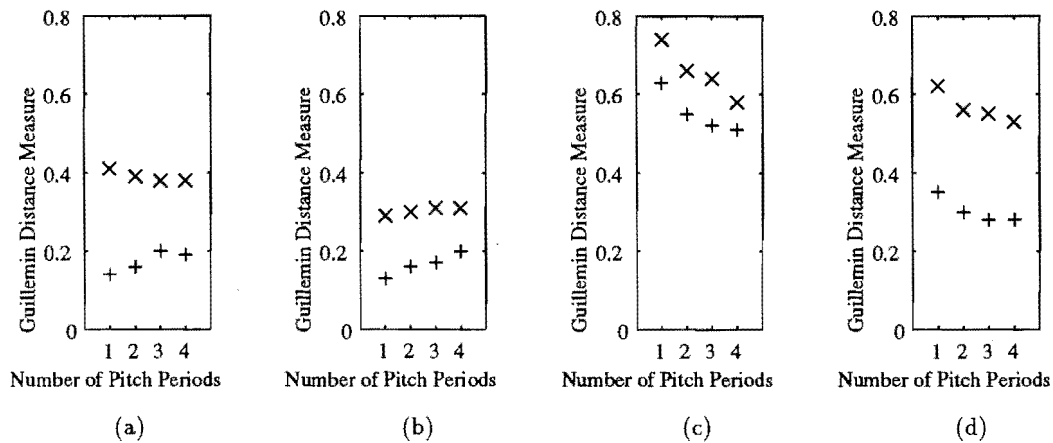


Figure 8.16. The distance measures resulting from the vocal tract reconstruction done from the four 200 Hz speech signals extracted from the original speech signal at the first zero position to the right of the main peak in the pitch period (indicated by 'x') and at the last zero position to the left of the main peak in the pitch period (indicated by '+') over four, three, two and one pitch periods of the signal for: (a) a 200 Hz signal made from filter 1 convolved with sawtooth 1, (b) a 200 Hz signal made from filter 1 convolved with sawtooth 2, (c) a 200 Hz signal made from filter 2 convolved with sawtooth 1, and finally (d) a 200 Hz signal made from filter 2 convolved with sawtooth 2.

8.4.6 Discussion About The New Method Of Glottal Pulse Removal

It is clear from the preliminary experiments on synthetic speech that if the glottal pulse is known it is possible to deconvolve it from the speech signal using a Wiener filter leaving the vocal tract filter response. This suggests it could be done for real speech if the effect that the radiation from lips has on the speech signal has already been accounted for. Once the filter response has been recovered, it is possible to reconstruct the vocal tract shape.

The preliminary results suggest that the number of pitch periods of the speech signal and excitation signal used in the Wiener filter affects the accuracy of the recovered areas. However the biggest influence on the accuracy is the position of the point of zero amplitude, relative to the major peak in the pitch period of the signal, from which the speech signal is extracted. The results suggest for sustained vowel sounds that the reconstruction process should be done pitch synchronously otherwise the vocal tract shape areas recovered from the vocal tract filter will vary.

The number of pitch periods of the excitation signal from which the Wiener filter was made, was the same number as the speech signal the Wiener filter was convolved with. The resulting signal was, as illustrated in 8.13 (j), the recovered filter response and some artefacts of the convolution. It would be worthwhile investigating what would happen to the accuracy of the recovered areas if the Wiener filter were made from only one pitch period. The result of convolving the Wiener filter made from one pitch period of the excitation with the speech signal would be a train of filter responses (with a pitch period the same as the speech signal) plus the artefacts of convolution. It should then be possible to extract one 'pitch period' of the filter response train to calculate the recovered areas. If this were possible it would mean the accuracy of the recovered areas would not be dependent on "the position of the point of zero amplitude, relative to the major peak in the pitch period of the signal, from which the speech signal is extracted", since one could always ensure that the filter response, extracted from the filter response train, would always give the most accurate area recovery.

To date the testing of the glottal pulse deconvolution method has been carried out in ideal conditions: no noise, a known vocal tract filter, and excitation signal. It is important to learn what effect noise will have on the reconstruction process. In addition it would be useful to repeat the experiments using an excitation signal made from trying to model the actual glottal pulse, such as Rosenberg's model (Rosenburg, 1971). Whilst this is still not the real glottal pulse it would be useful to establish whether it could be used in the Wiener filtering method of deconvolution. It was established, in sec 8.4.3, that the Wiener filtering works better for some signals than others.

The results in sec 8.4.3 suggest that it is possible to deconvolve a different excitation signal from the speech signal than the original one used in the speech signal and still get reasonable vocal tract shape reconstructions. This suggests that it may not be necessary to know the exact shape of the glottal pulse to recover the vocal tract shape. However much more must be learnt about the nature of the true glottal pulse before this can be established. In addition the synthetic speech used thus far in the investigation did not include a lip radiation component. In order to establish that the deconvolution process will work for real speech, the effect of lip radiation on the speech signal must be correctly modelled and then removed from the speech signal before the glottal pulse is deconvolved. Despite the large number of unknowns, it does seem that the proposed

new method of glottal pulse deconvolution is worth investigating further.

CHAPTER 9

CONCLUSIONS

The visual representation of the features of speech and its application to speech therapy is the major topic of the work presented in this thesis. The second main topic is the requirements for the development of an effective visual speech aid. A consequence of the work is the visual speech aid, the CASTT. However the most significant contribution of the work presented in this thesis is the VDT.

Speech has many features which can be calculated from the acoustic signal: loudness, pitch, duration, quality, to name a few (see sections 1.4 and 2.1). For the speech-impaired, transforming speech from an acoustic signal to a visual signal can provide feedback, additional to aural feedback, about the speech features of an utterance. Over the last 120 years various visual speech aids have been developed (see sections 2.2 and 2.3). None, however, have had the same combination of speech analysis modules as the CASTT.

The CASTT is a real-time computer-based visual speech aid, which consists of 8 speech analysis modules. These are: the Loudness Monitor, the Voice Pitch Tracker, the Concurrent Pitch and Loudness module, the Spectrogram module, the Sustained Phonation module, the Fricative Monitor, the Vocal Tract Shape module and the Lisajous Figure module (the speech analysis algorithms utilised in the speech modules of the CASTT are given in Chapter 4). The aid has been designed to remediate errors in suprasegmental aspects of speech, in articulation and in phonation (see sec 3.2).

The ability to visually display various speech features does not, itself, guarantee a speech aid's effectiveness. The aid must be able to display the features of speech in real-time. In addition it must be easy for the therapist and client to use, it must impart information which is of use in speech therapy and it must fit the requirements of speech therapy. These, for the most part, can be achieved through developing the aid interactively with speech therapists (see Chapter 5). The final test to gauge the effectiveness of a visual speech aid is the VDT.

It is the premise of this thesis that an effective visual speech aid must provide a visual response from which a judgement on the "goodness" of an utterance can be made. The VDT is a two-part test which assesses a visual speech aid's capacity to do this. It is proposed that two things are required from a speech aid for it to be effective. Firstly, the clusters of acceptable utterances must be separate from the clusters of unacceptable utterances in display space (see sec 6.4). Secondly, the acoustic features which distinguish acceptable utterances from unacceptable utterances must be evident in the displays of the speech aid (see sec 6.5). Two separate tests are required to assess these requirements, VDT part I and VDT part II (see sections 6.4 and 6.5).

The CASTT was assessed by 15 speech therapists and through the VDT. Improve-

ments can still be made to all its modules (see sections 5.2.1, 6.5.1.1, 6.5.1.2, 5.1.1.1 and 6.5.1.3). The aid is weak for errors in fricatives and voiced/unvoiced distinction (see sec 6.5.3). Two of the modules, the Fricative Monitor and Vocal Tract Shape module, require extensive changes (see sections 5.2.1 and 8.3.4.1). However, even if no more improvements or changes to the CASTT are made, the CASTT is a speech aid which can be used to remediate many speech errors (see sections 5.2.1 and 6.5.3).

The work presented in this thesis has also made contributions to methods of speech analysis. A recognition algorithm for isolated fricative sounds of English, based on spectral classification was developed (see sec 7.6). This algorithm has potential applications in a speech analysis module in a visual speech aid (see sec 7.7). Finally a new two-point vocal tract reconstruction method is proposed. The Direct Estimation method, utilised in the Vocal Tract Shape module in the CASTT, was rejected because the reconstructions of New Zealand open and back vowels, [o,ɔ,ɒ,u,a], were inaccurate and inconsistent (see sec 8.3.4.1). In the Direct Estimation method the glottal pulse is removed indirectly from the speech signal through flattening the spectrum. The new method involves removing the glottal pulse directly from the speech signal by deconvolving the measured glottal pulse from the signal using a Wiener filter. Reconstructing the vocal tract shape from the resulting vocal tract filter signal is then performed in the same manner as in the Direct Estimation method.

The idea has been tested out on synthetic speech with known vocal tract shapes and excitation signals (see sec 8.4). Some excitation signals make better Wiener filters than others, therefore it is necessary to test whether a Wiener filter can be made from a measured glottal pulse (see sec 8.4.6). The accuracy of the reconstructed vocal tract shapes is strongly influenced by the position in the pitch cycle (the beginning of the pitch cycle is taken to be the point with the maximum amplitude) from which the signal was extracted (see sec 8.4.4). The least accurate reconstruction always occurs if the signal is extracted near or at the beginning of the pitch cycle (see sections 8.4.4 and 8.4.5). Whilst the proposed two-point method requires much more investigation (see sec 8.4.6) it appears to have some promise as a potential method of reconstructing the vocal tract shape from the speech signal.

A substantial part of the first half of this thesis work was devoted to consultation with the speech therapy community and to co-ordinating their contributions to the project. The second half focussed, primarily, on the VDT. The work on vocal tract reconstruction was on-going throughout the entire time.

A series of minor contributions were made in the work of this thesis, in addition to the major ones outlined above. All the analysis algorithms utilized in the CASTT were formalized and corrected (see sections 4.3, 4.4, 4.5.3, 4.6.1, and 4.7.1 and figure 5.1). The display formats of the Loudness Monitor, the Voice Pitch Tracker and the Vocal Tract Shape modules were changed substantially and new analysis options were added to each of the modules (see sections 4.2.2 and 5.1.1.1). The TMS32010 code for the Vocal Tract Shape module, also, was changed substantially (see sec 4.2.2). The IBM-PC code was written for the Concurrent Loudness and Pitch module (see sec 4.2.2), and the analysis algorithm for Lissajous figures was devised (see sec 4.8). Finally, the speech list, which represents the common speech errors, proposed by Braeges and Houde (1982) was altered. Two changes were made. Firstly, the list was adapted for New Zealand English (only one change was required, see sec 6.3.1). Secondly, a smaller list was compiled to test modules with current-value-plots; the first list was designed

to test modules with time-plots only (see sec 6.3.1).

To conclude, transforming acoustic features of speech into visual responses can have powerful applications in speech therapy. However to guarantee the effectiveness of a visual speech aid, it must be developed interactively with speech therapists, the speech analysis algorithms utilized by the aid must be accurate or at least predictable, and the aid itself must be tested thoroughly with the VDT.

APPENDIX A

QUESTIONNAIRES

A.1 THE QUESTIONNAIRE FOR THE FIRST ASSESSMENT PERIOD

The following five questions were asked for each of the CASTT modules:

- What did you think of the visual display and the method of presentation ?
- Was there any ambiguity in the presentation of the module ?
- Can you see a use or potential for this module ?
- Did you have any particular likes or dislikes about the module ?
- Are there any areas you think need improvement ?

The following were general questions asked about the CASTT.

- What age group do you think would benefit most from a computer package like this ?
- Do you think incorporating games into the CASTT would be a good idea ?
- Can you see a need for other modules ?, if so what ?
- Would you use the CASTT yourself ?

A.2 THE QUESTIONNAIRE FOR THE SECOND ASSESSMENT PERIOD

The following four questions were asked about each of the CASTT modules:

- Can you see any use for this program ?
- Can you see any potential for this program ?, what is it and what changes need to be done to achieve this ?
- What did you not like about this program ?
- Have you any other suggestions ?

The rest of the questions were specifically directed towards each module.

Loudness Monitor

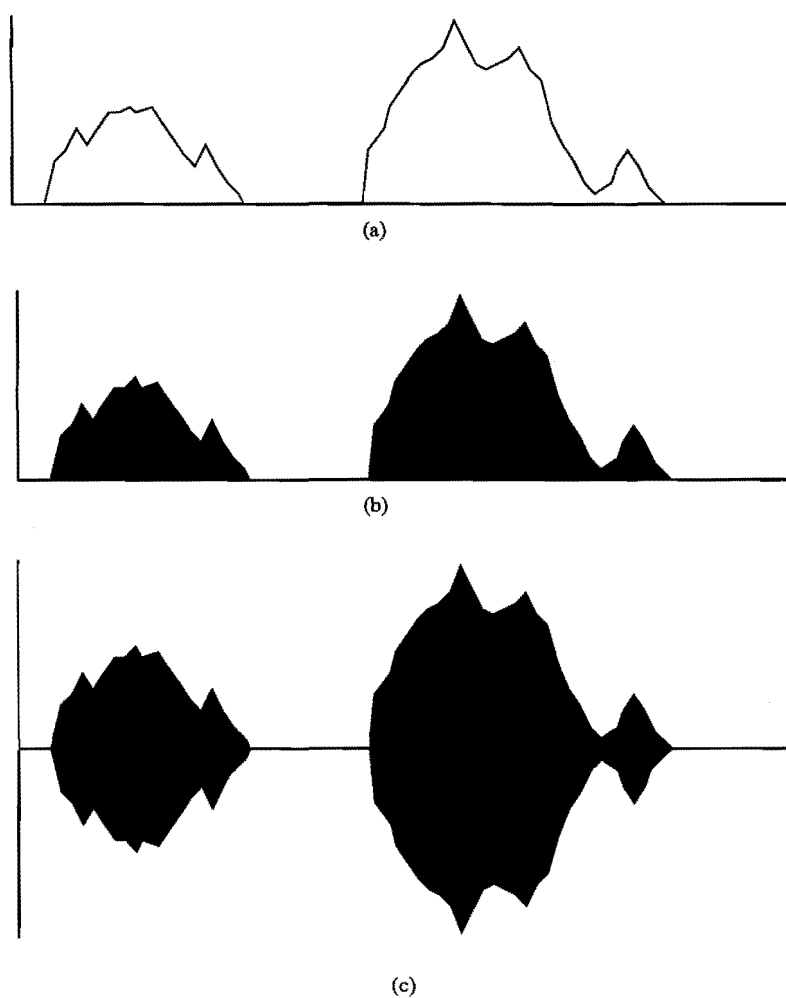


Figure A.1. The Loudness Information Display Options a/ current display, b/ blocked in contours, c/ reflected blocked contours.

- Is the visual display easy to understand ?
- Would you like the display to be plotted as [b/ in fig A.1] or as [c/ in fig A.1]?
- Is it necessary to have the overlay and superimpose function?, Would you get rid of one?, if so which ?
- Is the line option useful? or just a hindrance on the screen ?
- Currently 3 sec is the time scale default (the length of time the loudness wave form is plotted on the screen), you are able to change it to 6 or 9 secs. Would you like the default to be different if so what (you have the choice of 6 or 9 secs) ?
- Is the option which moves the vertical axis to the right useful ?
- Do you use the option that plays back the last 3 secs of inputted speech? Have you any comments ?
- Do you use the expand option which expands a selected 3 second segment of a 6 or 9 sec plotted display ?
- Do you use the scale option? is it necessary ?
- Do you use the disk option?, is it easy to use?, is more explanation about it necessary? Is the option necessary ?

Voice Pitch Tracker Module

- Is the visual display easy to understand ?
- Is it necessary to have the overlay and superimpose function?, Would you get rid of one?, if so which ?
- Currently 3 sec is the time scale default (the length of time the loudness wave form is plotted on the screen), you are able to change it to 6 or 9 secs. Would you like the default to be different if so what (you have the choice of 6 or 9 secs) ?
- Is the option which moves the vertical axis to the right useful ?
- Do you use the option that plays back the last 3 secs of inputted speech ? Have you any comments ?
- Do you use the expand option which expands a selected 3 second segment of a 6 or 9 sec plotted display ?
- Do you use the disk option?, is it easy to use?, is more explanation about it necessary? Is the option necessary ?
- What do you think about the data page?, does it convey useful information?, do any of the values seem wrong to you? Is there any thing else that you would like the data page to display ?

Concurrent Loudness And Pitch

- Do you find this program useful ?



Figure A.2. An alternative way of displaying the concurrent pitch and loudness module contours.

- Would you prefer both plots to be on the same display graph i.e. [see figure A.2] ?
- Would you like to see more analysis options (eg. superimpose, save to disk etc) ? if so what options ?
- Would you like to be able to change the time scale ? (like in pitch analysis and loudness analysis)

Spectrogram Module

- Do you understand how the display is representing the spectrogram? Do you understand what the colours mean ?
- Is there too much writing on the screen ?
- How much of the spectral information do you want to look at (e.g. just to 3 khz say) ?
- What is the most useful part of the spectrum? (e.g. the 1st and 2nd formant ?) Would you prefer the spectral information to be displayed in a different way ? if so have you any suggestions (e.g. having a display that plots the frequencies of the first and second formants)?
- Would you like the option to change the time scale, like the pitch analysis program ?

Fricative Monitor

- Is this program a consistent fricative monitor ?
- Does this program register fricatives better for some people than others? if so who ?
- Is the way the fricative information is displayed easy to understand ?
- Would the fricative display be more effective if it was represented as a meter ?
- Do you understand about setting the level at the beginning of the program?
- Would you like more information on it ?

Vocal Tract Shape Module

- Is it useful having the 2 displays ?
- Is the overlay option useful ?
- Would you like the option of displaying one head or two heads ?
- Do you think the program is consistent and/or reliable ?
- Would you find the program more useful if the tongue position was included ?

General Comments About the CASTT

- What age group do you think would benefit most from a computer package like the CASTT ?
- Do you think incorporating games, either for leisure or learning purposes into the CASTT would be a good idea?
- Can you suggest some ideas for games (eg. a balloon that gets bigger and bigger as sounds spoken in to the microphone get louder and louder)?
- Have you any suggestions for other analysis programs (eg. a nasal monitor)?
- Have you any suggestions for ways of displaying the information in the analysis program/s you suggested?
- Have you any suggestions for new ways of displaying the information in the current analysis methods?
- Would you use CASTT yourself?
- Have you any suggestions of brands of computers the CASTT, other than the IBM-PC, should be able to run on (eg. Apple IIe)?
- If the CASTT came on the market would you buy it ?
- Do you find the CASTT easy to use ?
- Could it be made easier? (any suggestions ?)
- Is the on-line help easy to understand? is it useful? should more be written ?
- Do you find the time taken to load the programs was too long ?
- Have you any other comments or queries ?

APPENDIX B

THE VISUAL DISPLAY TEST

B.1 DATA FOR PRELIMINARY VDT

| Successful Identification | NZ-SL1 | | NZ-SL2 | |
|-----------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| | 8up level | 7up level | 8up level | 7up level |
| all four speech pairs | 4.22×10^{-6} | 1.89×10^{-3} | 1.75×10^{-6} | 7.82×10^{-4} |
| both different-speech pairs | 1.1×10^{-2} | 2.34×10^{-1} | 4.58×10^{-3} | 9.69×10^{-2} |
| both same-speech pairs | 1.1×10^{-2} | 2.34×10^{-1} | 4.58×10^{-3} | 9.69×10^{-2} |

Table B.1. The random probabilities of successful identification in the preliminary VDT.

| | Loudness | | | | Pitch | | | | Spectrogram | | | |
|-------------------------|----------|----------|----------|----------|----------|----------|----------|----------|-------------|----------|----------|----------|
| | x_1x_2 | y_1y_2 | x_1y_1 | x_2y_2 | x_1x_2 | y_1y_2 | x_1y_1 | x_2y_2 | x_1x_2 | y_1y_2 | x_1y_1 | x_2y_2 |
| boo/boot | 7 | 1 | 9 | 3 | 2 | 1 | 8 | 6 | 7 | 7 | 9 | 9 |
| allay/away | 7 | 5 | 9 | 9 | 4 | 9 | 9 | 9 | 5 | 7 | 5 | 3 |
| two/do | 6 | 7 | 7 | 7 | 5 | 7 | 9 | 9 | 8 | 6 | 5 | 4 |
| poppa/pop | 7 | 8 | 4 | 5 | 9 | 9 | 9 | 8 | 7 | 5 | 8 | 9 |
| ban/ran | 7 | 8 | 1 | 3 | 0 | 0 | 6 | 4 | 3 | 4 | 7 | 9 |
| I/shy | 5 | 2 | 5 | 9 | 1 | 4 | 9 | 6 | 8 | 5 | 6 | 9 |
| oh/toe | 3 | 1 | 8 | 6 | 3 | 4 | 2 | 0 | 7 | 7 | 8 | 9 |
| Ann/van | 4 | 0 | 1 | 9 | 8 | 3 | 9 | 6 | 6 | 0 | 8 | 3 |
| Sue/zoo | 5 | 6 | 8 | 2 | 1 | 2 | 6 | 3 | 4 | 8 | 1 | 0 |
| me/bee | 9 | 8 | 8 | 9 | 0 | 1 | 8 | 9 | 6 | 8 | 7 | 5 |
| me/mbee | 9 | 6 | 9 | 9 | 0 | 4 | 1 | 9 | 7 | 1 | 5 | 8 |
| see/she | 0 | 0 | 9 | 5 | 2 | 1 | 8 | 4 | 3 | 6 | 9 | 8 |
| be/ba | 8 | 7 | 6 | 3 | 1 | 1 | 8 | 9 | 8 | 3 | 7 | 5 |
| do/door | 7 | 1 | 6 | 2 | 7 | 3 | 7 | 9 | 8 | 1 | 7 | 5 |
| Ed/add | 8 | 4 | 5 | 7 | 9 | 9 | 7 | 9 | 4 | 9 | 9 | 9 |
| fought/foot | 1 | 2 | 7 | 6 | 0 | 1 | 9 | 8 | 8 | 3 | 8 | 5 |
| down/don | 5 | 0 | 6 | 9 | 0 | 7 | 9 | 9 | 6 | 5 | 8 | 8 |
| use/ooze | 4 | 1 | 6 | 5 | 0 | 5 | 4 | 5 | 4 | 5 | 4 | 7 |
| way/ray | 4 | 5 | 3 | 3 | 0 | 1 | 9 | 7 | 8 | 8 | 9 | 4 |
| spot/spot | 8 | 0 | 3 | 9 | 1 | 1 | 7 | 7 | 0 | 1 | 5 | 9 |
| now?/now! | 9 | 8 | 9 | 8 | 8 | 1 | 9 | 9 | 8 | 8 | 5 | 8 |
| contract(n)/contract(v) | 3 | 8 | 8 | 6 | 6 | 7 | 6 | 9 | 4 | 8 | 6 | 3 |
| normal/high pitch | 0 | 8 | 8 | 5 | 3 | 9 | 6 | 6 | 7 | 9 | 5 | 2 |
| normal/loud | 0 | 8 | 9 | 4 | 3 | 2 | 9 | 5 | 7 | 9 | 9 | 8 |
| normal/soft | 0 | 6 | 5 | 9 | 3 | 5 | 3 | 6 | 7 | 6 | 7 | 8 |
| normal/slow | 0 | 9 | 7 | 9 | 3 | 5 | 9 | 8 | 7 | 6 | 8 | 9 |
| normal/nasal | 0 | 6 | 9 | 7 | 3 | 3 | 6 | 6 | 7 | 5 | 7 | 8 |
| normal/breathy | 0 | 9 | 9 | 9 | 3 | 5 | 8 | 4 | 7 | 9 | 9 | 9 |
| normal/tense | 0 | 9 | 9 | 8 | 3 | 9 | 8 | 9 | 7 | 9 | 9 | 9 |

Table B.2. The raw data obtained for the time-plots in the preliminary VDT. Each column shows the number correctly identified speech pair types, out of a possible 9 (X_1X_2, Y_1Y_2 being same-speech pairs, and $(X_1Y_1, X_2Y_2$ being different-speech pairs).

B.2 PROBABILITIES USED IN THE VDT

The participants were presented with the plots of three utterances, two formed a same-speech pair and the third was an error utterance. The probability of a participant successfully identifying a same-speech pair in a plot-set, through random guessing, is 0.33. The probability of x participants successfully identifying the same-speech pairs in the four plot-sets associated with each elementary error, through random guessing is:

$$\left(\sum_{i=x}^{31} \frac{31!}{i!(31-i)!} 0.33^i (0.67)^{31-i} \right)^4 \quad (\text{B.1})$$

where $x=28$ if it is for random guessing at the 90+% level and $x=25$ if it is for random guessing at the 80+% level.

| | Vocal Tract Shape | | | | Fricative Content Response | | | | Lissajous Figures | | | |
|---------|-------------------|----------|----------|----------|----------------------------|----------|----------|----------|-------------------|----------|----------|----------|
| | x_1x_2 | y_1y_2 | x_1y_1 | x_2y_2 | x_1x_2 | y_1y_2 | x_1y_1 | x_2y_2 | x_1x_2 | y_1y_2 | x_1y_1 | x_2y_2 |
| [t]/[d] | 8 | 0 | 7 | 7 | 9 | 1 | 0 | 8 | 0 | 6 | 9 | 9 |
| [i]/[ɪ] | 3 | 2 | 9 | 9 | 9 | 8 | 9 | 0 | 8 | 9 | 9 | 9 |
| [æ]/[v] | 3 | 8 | 9 | 8 | 8 | 1 | 8 | 8 | 3 | 7 | 9 | 8 |
| [s]/[z] | 0 | 9 | 0 | 9 | 3 | 2 | 8 | 8 | 7 | 2 | 9 | 9 |
| [m]/[b] | 8 | 5 | 9 | 9 | 9 | 5 | 9 | 9 | 9 | 4 | 9 | 9 |
| [ʃ]/[ʒ] | 0 | 2 | 0 | 7 | 3 | 8 | 0 | 9 | 7 | 9 | 0 | 9 |
| [i]/[a] | 3 | 9 | 9 | 8 | 9 | 8 | 9 | 9 | 8 | 9 | 5 | 1 |
| [u]/[ʊ] | 0 | 8 | 2 | 5 | 9 | 9 | 1 | 2 | 0 | 2 | 9 | 0 |
| [e]/[æ] | 9 | 3 | 8 | 8 | 9 | 8 | 6 | 7 | 1 | 3 | 6 | 9 |
| [ɔ]/[ʊ] | 8 | 7 | 6 | 7 | 9 | 9 | 9 | 0 | 2 | 3 | 4 | 6 |
| [j]/[u] | 7 | 0 | 7 | 3 | 6 | 9 | 9 | 9 | 8 | 0 | 9 | 7 |
| [w]/[r] | 7 | 5 | 2 | 4 | 0 | 9 | 9 | 0 | 1 | 9 | 9 | 6 |

Table B.3. The raw data obtained for the current-value-plots in the preliminary VDT. Each column shows the number correctly identified speech pair types, out of a possible 9 (X_1X_2, Y_1Y_2 being same-speech pairs, and $(X_1Y_1, X_2Y_2$ being different-speech pairs).

REFERENCES

- ABDELHAMIED, K., HASHISH, M., EMAN, O., KAMAL, D., EL-HETW, A. and AHMED, K. (1987), 'A multi-feature computer-based speech-training system for deaf children', In *IEEE/ Ninth Annual Conference of the Engineering in Medicine and Biology Society*, pp. 1-2.
- ADAMS, F.R., CREPY, H., JAMESON, D. and THATCHER, J. (1989), 'IBM products for persons with disabilities', In *IEEE Global Telecommunications Conference*, Dallas, U.S.A., pp. 980-984.
- AGUILERA, S., BORRAJO, A., PARDO, J.M. and MUNOZ, E. (1986), 'Speech-analysis-based devices for diagnosis and education of speech and hearing impaired people', In *Int. Conf. on acoustics, speech and signal processing*, IEEE, Tokyo, pp. 641-644.
- AINSWORTH, W. (1988), *Speech Recognition By Machine*, Peter Peregrinus Ltd.
- ARENDS, N., POVEL, D., VAN OS, E., MICHELSEN, S., CLAASSEN, J. and FEITER, I. (1991), 'An evaluation of the visual speech apparatus', *Speech Communication*, Vol. 10, No. 4, Nov., pp. 405-414.
- BADIN, P. (1989), 'Acoustics of voiceless fricatives: Production theory and data', *Speech Transmission Laboratory Quarterly Progress And Status Report*, No. 3/1989, pp. 33-55.
- BADIN, P. (1991), 'Fricative consonants: acoustic and x-ray measurements', *Journal of Phonetics*, Vol. 19, pp. 397-408.
- BATE, E.M., FALLSIDE, F., GULIAN, E., HINDS, P. and KEILLER, C. (1982), 'A speech training aid for the deaf with display of voicing, frication, and silence', In *Int. Conf. on acoustics, speech and signal processing*, IEEE, pp. 743-746.
- BATES, R.H.T. and MCDONNELL, M.J. (1986), *Image Restoration and Reconstruction*, Oxford University Press, Oxford.
- BATES, R.H.T., BRIESEMANN, N.P., CLARK, T.M., ELDER, A.G., FRIGHT, W.R., GARDEN, K.L., KENNEDY, W.K., SQUIRES, P.L., THORPE, C.W., TURNER, S.G. and JELINEK, H.J. (1987), 'Interactive speech-defect diagnostic/therapeutic /prosthetic aid', In LETELLIER, J.P. (Ed.), *Real Time Signal Processing X*, 20-21 August, pp. 131-139.
- BAUM, S.R. and BLUMSTEIN, S.E. (1987), 'Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in english', *Journal Of The Acoustical Society Of America*, Vol. 82, No. 3, pp. 1073-1077.

- BEHRENS, S.J. and BLUMSTEIN, S.E. (1988), 'Acoustic characteristics of english voiceless fricatives: A descriptive analysis', *Journal Of Phonetics*, Vol. 16, pp. 295-298.
- BERNSTEIN, L.E., FERGUSON, J.B. and GOLDSTEIN, M.H. (1986), 'Speech training aids for profoundly deaf children', In *Int. Conf. on Acoustics, Speech and Signal Processing*, IEEE, Tokyo, pp. 633-636.
- BERNSTEIN, L., GOLDSTEIN, M.H. and MASHIE, J.J. (1988), 'Speech training aids for hearing-impaired individuals: I. overview and aims', *Journal of Rehabilitation and Development*, Vol. 25, No. 4, pp. 53-62.
- BOOTHROYD, A. (1992), 'Impact of technology on the management of deafness', *The Volta Review*, Vol. 92, No. 4, May, pp. 74-82.
- BOOTHROYD, A., ARCHAMBAULT, P., ADAMS, R. and STORM, R.D. (1975), 'Use of a computer-based system of speech analysis and display in a remedial speech program for deaf children', *Volta Rev.*, Vol. 77, pp. 178-193.
- BORRILD, K. (1968), 'Experience with the design and use of technical aids for the training of deaf and hard of hearing children', *American Annals Of The Deaf*, Vol. 113, pp. 168-177.
- BRACEWELL, R.N. (1978), *The Fourier Transform and Its Applications*, McGraw-Hill Book Company, U.S.A.
- BRAEGES, J.L. and HOUDE, R.A. (1982), 'Use of speech training aids.', In SIMS, D., WALTER, G.G. and WHITEHEAD, R.L. (Eds.), *Deafness and Communication: Assessment and Training*, Williams and Wilkins, Baltimore, MD, pp. 222-244.
- BRIESEMANN, N. (1984), *A New Algorithm For Musical Pitch Estimation*, Master's Thesis, University of Canterbury, New Zealand.
- BRIGHAM, E. (1974), *The Fast Fourier Transform*, Prentice-Hall, Inc.
- BRISTOW, G.J. and FALLSIDE, F. (1978), 'Computer display of vowel quadrilateral with real-time representation of articulation', *Electronic Letters*, Vol. 14, No. 4, Feb, pp. 107-109.
- BRISTOW, G., FALLSIDE, F., GULIAN, E. and HINDS, P. (1981), 'Teaching frication to the hearing-impaired with the fricative training aid: A case study', *Journal of the British Association Teachers of the Deaf*, Vol. 5, No. 6, Nov., pp. 178-186.
- BROOKS, S. and FALLSIDE, F. (1976), 'A technique for converting the linear prediction areas model of speech to a simple articulatory model', In *Int Conf. on acoustics, speech and signal processing*, IEEE, Philadelphia, pp. 71-74.
- BROOKS, S., FALLSIDE, F., GULIAN, E. and HINDS, P. (1981), 'Teaching vowel articulation with the computer vowel trainer - methodology and results', *British Journal Of Audiology*, Vol. 15, pp. 151-163.
- BURRUS, C.S. and PARKS, T.W. (1985), *DFT/FFT and Convolution Algorithms, Theory And Implementation*, John Wiley and Sons, U.S.A.

- CHOU, L.M. (1985), *Pitch Estimation of Speech*, Final Year Project Report, University of Canterbury, New Zealand.
- CLARK, J. and YALLOP, C. (1990), *An Introduction To Phonetics And Phonology*, Basil Blackwell.
- COHEN, M. (1968), 'The ADL sustained phoneme analyzer', *American Annals Of The Deaf*, Vol. 113, pp. 248-252.
- CRICHTON, R.C. and FALLSIDE, F. (1974), 'Linear prediction model of speech production with applications to deaf speech training', *Proc. IEE*, Vol. 121, No. 8, pp. 865-873.
- CRYSTAL, D. (1980), *Introduction To Language Pathology*, Univerity Park Press, Baltimore.
- DE MANRIQUE, A. and MASSONE, M. (1981), 'Acoustic analysis and perception of spanish fricative consonants', *J. Acoustic Society of America*, Vol. 69, No. 4, April, pp. 1145-1153.
- DENOIX, B. (1984), 'A speech input and processing board for a personal computer', In *International Conference On Acoustic Signals And Speech Processing '84*, pp. 34B.4.1-34B.4.4.
- DESCOUT, R., TOUSIGNANT, B. and LECOURS, M. (1976), 'Vocal tract area function measurements: Two time-domain methods', In *IEEE International Conference On Acoustic Signals And Speech Processing '76*, Philadelphia, pp. 75-78.
- EL-BEZE, M. (1986), 'Refutation based recognition to help vowel articulation', *International Conference On Acoustic Signals And Speech Processing '86*, pp. 13.9.1-13.9.4.
- ELDER, A.G., BATES, R.H.T., BRIESEMANN, N.P., CLARK, T.M., FRIGHT, W.R., GARDEN, K.L., KENNEDY, W.K., SQUIRES, P.L., TURNER, S.G. and THORPE, C.W. (1987), 'Real time speech therapy aid', *New Zealand National Electronics Conference*, Vol. 24, pp. 115-118.
- FALLSIDE, F. and BROOKS, S. (1976), 'Real-time areagraph of continuous speech for analysis and speech training', *Electronic Letters*, Vol. 12, No. 20, Sept, pp. 515-516.
- FALLSIDE, F. and BROOKS, S. (1977), 'Some displays for computer-analysed speech', *Proc. IEE*, Vol. 124, No. 12, pp. 1227-1229.
- FERGUSON, J.B., BERNSTEIN, L.E. and GOLDSTEIN, M.H. (1988), 'Speech training aids for hearing-impaired individuals: II. configuration of the john hopkins aid', *Journal of Rehabilitation and Development*, Vol. 25, No. 4, pp. 63-68.
- FRY, D.B. (1979), *The Physics of Speech*, Cambridge University Press.
- GUILLEMIN, B.J. (1986), *Determining Vocal Tract Shape At High Pitch Using Linear prediction*, PhD thesis, University Of Auckland.

- GULIAN, E., FALLSIDE, F. and HINDS, P. (1984), 'Can deaf children interact with computers? evidence from speech acquisition training', In *Human-Computer Interaction - Interact '84, Proceedings of the IFIP Conference*, pp. 809-813.
- HARDCASTLE, W.J. (1976), *Physiology Of Speech Production - An Introduction For Speech Scientists*, Academic Press.
- HARRINGTON, J. (1988), 'Automatic recognition of english consonants', In JACK, M. and LAVER, J. (Eds.), *Aspects Of Speech Technology*, Edinburgh University Press, pp. 69-143.
- HAWKINS, P. (1984), *Introducing Phonology*, Hutchinson.
- HESS, W. (1992), 'Pitch and voicing determination', In FURUI, S. and SONDHJ, M.M. (Eds.), *Advances in Speech Signal Processing*, Marcel Dekker Inc, pp. 4-31.
- HUGHES, G.W. and HALLE, M. (1956), 'Spectral properties of fricative consonants', *Journal Of the Acoustical Society Of America*, Vol. 28, No. 2, Mar., pp. 303 - 310.
- HWANG, K. (1979), *Computer Arithmetic Principles, Architecture, And Design*, John Wiley and Sons, U.S.A.
- IBM (1988), 'Speech viewer (advertising pamphlet)'.
- INTEL (1993), 'Microprocessors vol.1', Intel Corporation.
- ITO, M.R. and ROBERTSON, R.W. (1971), 'Zero-crossing measurements for analysis and recognition of speech sounds', *IEEE trans AUDIO AND ELECTROACOUSTICS*, Vol. AU-19, No. 3, Sept., pp. 235-242.
- JAKOBSON, R. (1990), 'The concept of phoneme', In WAUGH, L.R. and MONVILLE-BURSTON, M. (Eds.), *On Language*, Harvard University Press, Chap. 15, pp. 217-241.
- JASSEM, W. (1965), 'The formant of fricative consonants', *Language And Speech*, Vol. 8, No. 3, pp. 1-16.
- JASSEM, W. (1979), 'Classification of fricative spectra using statistical discriminant functions', In LINDBLOM, B. and OHMAN, S. (Eds.), *Frontiers Of Speech Research*, Academic Press, pp. 77-91.
- JONGMAN, A. (1989), 'Duration of frication noise required for identification of english fricatives', *Journal Of The Acoustical Society Of America*, Vol. 85, No. 4, pp. 1718-1725.
- KALIKOW, D.N. and SWETS, J. (1972), 'Experiments with computer-controlled displays in second language learning', *IEEE Trans. Audio Electro-acoust.*, Vol. AU-20, mar, pp. 23-28.
- KAPLAN, H.M. (1971), *Anatomy and Physiology Of Speech*, McGraw and Hill Book Company.

- KEWLEY-PORT, D., WATSON, C.S. and CROMER, P. (1987a), 'The Indiana speech training aid (ISTRA): A microcomputer-based aid using speaker-dependent speech recognition.', In *American Speech-Hearing Language Foundation Computer Conference*, Houston, TX, pp. 94-97.
- KEWLEY-PORT, D., WATSON, C., MAKI, D. and REED, D. (1987b), 'Speaker-dependent speech recognition as the basis for a speech training aid', *International Conference On Acoustic Signals And Speech Processing '87*, April, pp. 372-375.
- KEWLEY-PORT, D., WATSON, C., ELBERT, M., MAKI, D. and REED, D. (1991), 'The indiana speech training aid (ISTRA) II: Training curriculum and selected case studies', *Clinical Linguistics and Phonetics*, Vol. 5, No. 1, pp. 13-38.
- LADEFOGED, P. (1972), *Elements Of Acoustic Phonetics*, The University Of Chicago Press.
- LADEFOGED, P. (1975), *A Course In Phonetics*, Harcourt Brace Jovanovich, Inc.
- LAVER, J. (1980), *The Phonetic Description Of Voice Quality*, Cambridge University Press.
- LEFEVRE, J., SOUMAGNE, J.C., TOUSIGNANT, B. and LECOURS, M. (1980), 'Mesures de la fonction d'aire du conduit vocal et estimation indirecte de la qualite des resultats', In *Actes de la conference IEEE sur les communications et l'energie*, Montreal, pp. 317-320.
- LEFEVRE, J., TOUSIGNANT, B. and LECOURS, M. (1981), 'Utilisation des methods acoustiques pour l'evaluation des fonctions d'aire du conduit vocal', In *12th Journees d'Etude sur la Parole, GALF*, Montreal, pp. 26-40.
- LEFEVRE, J.P., TOUSIGNANT, B. and LECOURS, M. (1983), 'Etude des configurations vocaliques des voyelles francaises a partir de mesures acoustiques', *Acustica*, Vol. 52, pp. 227-231.
- LEHISTE, I. (1976), 'Suprasegmental features of speech', In LASS, N.J. (Ed.), *Contemporary Issues In Experimental Phonetics*, Academic Press, Chap. 7, pp. 225-239.
- LEVELT, W.J.M. (1989), *Speaking - From Intention To Articulation*, MIT Press.
- LIEBERMAN, P. and BLUMSTEIN, S.E. (1988), *Speech Physiology, Speech Perception and Acoustic Phonetics*, Cambridge University Press.
- LIPPMANN, R.P. (1982), 'A review of research on speech training aids for the deaf', In LASS, N.J. (Ed.), *Speech and Language: Advances in Basic Research and Practice*, Academic Press, Inc, pp. 105-133.
- LUDEMAN, L. (1987), *Fundamentals Of Digital Signal Processing*, John Wiley and Sons, Inc.
- MAKHOUL, J. (1975), 'Linear prediction : A tutorial review', *Proc. IEEE*, Vol. 63, No. 4, April, pp. 561-580.

- MAKI, J.E., GUSTAFSON, M.S., CONKLIN, J.M. and HUMPHREY-WHITEHEAD, B.K. (1981), 'The speech spectrographic display: Interpretation of visual patterns by hearing-impaired adults', *Journal of Speech and Hearing Disorders*, Vol. 46, Nov, pp. 379-387.
- MARKEL, J.D. and Gray, Jr., A.H. (1976), *Linear prediction of speech*, Springer-Verlag, Berlin.
- MASHIE, J., HASEGAWA, A., HERBERT, E. and METZLER, M. (1985), 'A computerized approach to assessing and modifying the voice of deaf speakers', In *International Congress On Education Of The Deaf*, pp. 328-346.
- MASHIE, J., ALQUIST-VARI, D., WADDY-SMITH, B. and BERNSTEIN, L.E. (1988), 'Speech training aids for hearing-impaired individuals: III preliminary observations in the clinic and home', *Journal of Rehabilitation and Development*, Vol. 25, No. 4, pp. 69-82.
- MCKINNON, D. and LEE, H. (1976), 'Real-time recognition of unvoiced fricatives in continuous speech to aid the deaf', *International Conference On Acoustic Signals and Speech Processing*, pp. 586-589.
- MERMELSTEIN, P. (1967), 'Determination of the vocal tract shape from measured formant frequencies', *Journal of the Acoustic Society of America*, Vol. 41, No. 5, pp. 1283-1294.
- MILENKOVIC, P. (1984), 'Vocal tract area function from two-point acoustic measurements with formant frequency constraints', *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. assp - 32, No. 6, December, pp. 1122-1135.
- MILENKOVIC, P.H. and MULLER, E. (1985), 'Two point acoustic reconstruction of sustained vowel area functions', *Mathematics And Computers In Biomedical Applications*, pp. 351-362.
- MILLER, D.C. (1934), *The Science of Musical Sounds*, The Macmillan Company.
- MILLER, G.A. and NICELY, P. (1955), 'An analysis of perceptual confusions among some english consonants', *Journal Of The Acoustical Society Of America*, Vol. 27, No. 2, Mar., pp. 338-352.
- MOLHO, L. (1976), 'Automatic acoustic-phonetic analysis of fricatives and plosives', *International Conference On Acoustic Signals and Speech Processing*, pp. 182-185.
- MORSE, P.M. (1948), *Vibration and Sound*, McGraw and Hill.
- NATION, J.E. and ARAM, D.M. (1977), *Diagnosis Of Speech and Language Disorders*, The C.V. Mosby Company.
- NICKERSON, R.S. and STEVENS, K.N. (1973), 'Teaching speech to the deaf: Can a computer help?', *IEEE Trans. on Audio and Electro-acoust.*, Vol. AU-21, No. 5, Oct, pp. 445-455.

- NICKERSON, R.S., KALIKOW, D.N. and STEVENS, K.N. (1976), 'Computer-aided speech training for the deaf', *Journal of speech and hearing disorders*, Vol. 41, pp. 120-132.
- OSBERGER, M.J., JOHNSTONE, A., SWARTS, E. and LEVITT, H. (1978), 'The evaluation of a model speech-training program for deaf children', *Journal of Communication Disorders*, Vol. 11, pp. 293-313.
- OSBERGER, M.J., MOELLER, M.P., KROESE, J.M. and LIPPMANN, R.P. (1981), 'Computer-assisted speech training for the hearing-impaired', *Journal of the Academy of Rehabilitation Audiology*, Vol. 14, No. 145-158, pp. 145-158.
- OSBERGER, M.J., LIPPMANN, R.P., MOELLER, M.P. and KROESE, J.M. (1982), 'Evaluation of a computer speech-training aid for the deaf', In RAVIV, J. (Ed.), *Use of computers in aiding the disabled*, North Holland, Amsterdam, pp. 231-247.
- OWENS, F. (1993), *Signal Processing Of Speech*, McGraw-Hill, Inc.
- PARDO, J.M. (1982), 'Vocal tract shape analysis for children', In *International Conference on Acoustics, Speech and Signal Processing*, IEEE, Paris, pp. 763-766.
- PENTZ, A., GILBERT, H.R. and ZAWADZKI, P. (1979), 'Spectral properties of fricative consonants in children', *Journal Of the Acoustical Society Of America*, Vol. 66, No. 6, Dec., pp. 1891-1893.
- PICKETT, J.M. (1972), 'Status of speech analyzing communication aids for the deaf', *IEEE Trans. On Audio And Electroacoustics*, Vol. AU-20, No. 1, March, pp. 3-8.
- PICKETT, J.M. and CONSTAM, A. (1968), 'A visual speech trainer with simplified indication of vowel spectrum', *American Annals of the Deaf*, Vol. 113, pp. 253-258.
- POVEL, D. (1974a), 'Development of a vowel corrector for the deaf', *Psychological Research*, Vol. 37, pp. 51-70.
- POVEL, D.J. (1974b), 'Evaluation of the vowel corrector as a speech training device for the deaf', *Psychological Research*, Vol. 37, pp. 71-80.
- POVEL, D. and ARENDS, N. (1991), 'The visual speech apparatus: Theoretical and practical aspects', *Speech Communications*, Vol. 10, No. 1, Feb., pp. 59-80.
- POVEL, D. and WANSINK, M. (1986), 'A computer-controlled vowel corrector for the hearing impaired', *Journal of Speech and Hearing Research*, Vol. 29, March, pp. 99-105.
- PRONOVOST, W. (1963), 'A pilot study of the voice visualizer for teaching speech to the deaf', In *Proceedings of the International Congress on Education of the Deaf*, pp. 925-930.
- PRONOVOST, W. (1967), 'Developments in visual displays of speech information', *Volta Rev.*, Vol. 69, pp. 365-373.

- PRONOVOST, W., YENKIN, L. and LERNER, R. (1968), 'The voice visualizer', *American Annals of the Deaf*, Vol. 113, pp. 230–238.
- RABINER, L.R. and SCHAFER, R.W. (1978), *Digital signal processing of speech signals*, Prentice-Hall, Englewood Cliffs, New Jersey 07632, USA.
- RISBERG, A. (1968), 'Visual aids for speech correction', *American Annals of the Deaf*, Vol. 113, pp. 178–194.
- ROBERTS, R.A. and MULLIS, C.T. (1987), *Digital Signal Processing*, Addison-Wesley Publishing Company.
- ROSENBERG, A.E. (1971), 'Effect of glottal pulse shape on the quality of natural vowels', *Journal of the Acoustic Society of America*, Vol. 49, pp. 583 – 590.
- SCHROEDER, M.R. (1967), 'Determination of the geometry of the human vocal tract by acoustic measurements', *Journal of the Acoustic Society of America*, Vol. 41, No. 4, pp. 1002 –1010.
- SEARSON, M. (1965), 'A speech-training programme using the kamplex visible speech apparatus', *Teacher of the Deaf*, Vol. 43, pp. 89–95.
- SHIGENAGA, M. and KUBO, H. (1986), 'Speech training systems for handicapped children using vocal tract lateral shapes', In *International Conference on Acoustics, Speech and Signal Processing*, IEEE, Tokyo, pp. 636–640.
- SHIGENAGA, M., TATSUMI, H., SONE, O. and SEKIGUCHI, Y. (1981), 'Some speech training systems for hearing impaired children', *Computers in Education*, North Holland, pp. 447–454.
- SKINNER, P.H. and SHELTON, R.L. (Eds.) (1978), *Speech, Language, and Hearing: Normal Processes and Disorders*, Addison-Wesley Publishing Company.
- Software Research 'Advertisement brochure on the visible speech aid'.
- SONDHI, M.M. (1979), 'Estimation of vocal tract areas: The need for acoustical measurements', *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP-27, No. 3, June, pp. 268 –273.
- SONDHI, M.M. (1984), 'A survey of the vocal tract inverse problem: Theory, computations and experiments', In SANROSA, F., PAO, Y., SYMES, W. and HOLLAND, C. (Eds.), *Inverse Problems Of Acoustic and Elastic Waves*, Society For Industrial and Applied Mathematics, Philadelphia, pp. 1–19.
- SONDHI, M.M. and GOPINATH, B. (1971), 'Determination of vocal tract shape from impulse response at the lips', *Journal of the Acoustic Society of America*, Vol. 49, No. 6, pp. 1867 –1873.
- SONDHI, M.M. and RESNICK, J.R. (1983), 'The inverse problem for the vocal tract: Numerical methods, acoustical experiments and speech synthesis', *Journal Of The Acoustical Society Of America*, Vol. 73, No. 3, March, pp. 985 –1002.

- SOUTHARD, J.R. (1983), 'Macpitts: An approach to silicon computation', *Computer*, Vol. 16, No. 12, Dec., pp. 74-82.
- STARK, R.E. (1971), 'The use of real-time visual displays of speech in the training of a profoundly deaf, nonspeaking child: A case report', *Journal Of Speech and Hearing Disorders*, Vol. 36, No. 3, pp. 397-409.
- STARK, R.E. (1972), 'Teaching features of speech to deaf children by means of real-time visual displays', In *Proc. Int. Symp. Speech Comm. and Profound Deafness*, A. G. Bell Assoc., Washington D.C., pp. 335-343.
- STARK, R.E., CULLEN, J.K. and CHASE, R.A. (1968), 'Preliminary work with the new bell telephone visible speech translator', *American Annals of the Deaf*, Vol. 113, pp. 205-214.
- STEVENS, K.N. (1971), 'Airflow and turbulence noise for fricative and stop consonants: Static considerations', *Journal Of The Acoustical Society Of America*, Vol. 50, No. 4, pp. 1180-1192.
- STEVENS, K.N., KALIKOW, D.N. and WILLEMAIN, T.R. (1975), 'A miniature accelerometer for detecting glottal waveforms and naslization', *Journal of Speech and Hearing Research*, Vol. 18, pp. 594-599.
- STEWART, L.C., LARKIN, W.B. and HOUDE, R.A. (1976), 'A real-time spectrograph with implications for speech training for the deaf', In *International Conference on Acoustics, Speech and Signal Processing*'76, IEEE, Philadelphia, pp. 590-593.
- STREVEN, P. (1960), 'Spectra of fricative noise in human speech', *Language and Speech*, Vol. 3, pp. 32-49.
- THOMAS, I.B. (1968), 'Real-time visual display of speech parameters', In *Proceedings Of The National Electronics Conference*, Chicago, pp. 382-387.
- THOMAS, I.B. and SNELL, R.C. (1970), 'Articulation training through visual speech patterns', *Volta Review*, Vol. 72, pp. 310-318.
- TOUSIGNANT, B., LEFEVRE, J. and LECOURS, M. (1979), 'Speech synthesis from vocal tract area function acoustical measurements', In *IEEE International Conference On Acoustic Signals And Speech Processing*'79, Washington D.C., pp. 921-924.
- TURNER, S.G. (1986), *Real-Time Speech Analysis for use with Impaired Speech Aids*, Master's Thesis, Electrical and Electronic Engineering, University of Canterbury, New Zealand.
- VEMULA, N.R., ENGBRETSON, A. and ELLIOTT, D. (1982), 'Estimation of vocal tract shape from input/output measurements', In *International Conference On Acoustic Signals And Speech Processing*, pp. 927-930.
- WAKITA, H. (1973), 'Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms', *IEEE Trans. Audio and Electroacoustics*, Vol. au-21, No. 5, October, pp. 417-427.

- WAKITA, H. (1979), 'Estimation of vocal tract shapes from acoustical analysis of the speech wave: The state of the art', *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP-27, No. 3, June, pp. 281-285.
- WATSON, C.S., KEWLEY-PORT, D., REED, D.J. and MAKI, D. (1989), 'The Indiana speech training aid (ISTRA) I: Comparisons between human and computer-based evaluation of speech quality', *Journal of Speech and Hearing Research*, Vol. 32, June, pp. 245-251.
- WITTEN, I.H. (1982), *Principles Of Computer Speech*, Academic Press (London).